12-2019

Statistical methods for estimating and testing treatment effect for multiple treatment groups in observational studies.

Xiaofang Yan
University of Louisville

STATISTICAL METHODS FOR ESTIMATING AND TESTING TREATMENT EFFECT FOR MULTIPLE TREATMENT GROUPS IN OBSERVATIONAL STUDIES

By

Xiaofang Yan
B.S., Xi'an Jiaotong University, 2011
M.S., Chinese Academy of Sciences, 2015

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

December 2019

STATISTICAL METHODS FOR ESTIMATING AND TESTING TREATMENT EFFECT FOR MULTIPLE TREATMENT GROUPS IN OBSERVATIONAL STUDIES

By

Xiaofang Yan
B.S., Xi'an Jiaotong University, 2011
M.S., Chinese Academy of Sciences, 2015

A Dissertation Approved on

November 20, 2019

by the following Dissertation Committee:

Dr. Maiying Kong, Dissertation Director

Dr. Bakeerathan Gunaratnam

Dr. Karunarathna B. Kulasekera

Dr. John Allen Myers

Dr. Qi Zheng

DEDICATION

This dissertation is dedicated to my parents who have given me invaluable educational opportunities. Their affection, love and encouragement laid a solid foundation in my heart.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Maiying Kong for her insightful guidance and constant inspiration for my research over the last four years. I also appreciate her encouragement and care when I had difficult time, in particular during the time I was in job market.

I would like to thank Drs. Karunarathna B. Kulasekera, Bakeerathan Gunaratnam, John Allen Myers and Qi Zheng for their time to serve as my dissertation committee. I especially would like to thank Dr. Zheng for his great input in my recent two projects on causal inference, and thank Dr. John Allen Myers for providing me the opportunity to work on Medicaid and Medicare data. I would like to thank the Department of Bioinformatics and Biostatistics for providing me a great environment to do research, thank all the faculty members for enriching my knowledge in biostatistics, and thank my fellow students in the department who have helped me and made my life more enjoyable during my study.

ABSTRACT

STATISTICAL METHODS FOR ESTIMATING AND TESTING AVERAGE TREATMENT EFFECT FOR MULTIPLE TREATMENT GROUPS IN OBSERVATIONAL STUDIES

Xiaofang Yan

November 20, 2019

This dissertation consists of three projects related to causal inference based on observational data.

In my first project, I conduct two investigations. The first one is to use rank aggregation technique to select the data-driven optimal propensity score estimation method from four existing methods, which include logistic regression model, covariate balancing propensity score, random forest and generalized boosted model. The optimal measure is their performance in balancing the covariates. The second investigation is to use the ensemble approach to improve the outcome estimation, and further improve the ATE estimation using the doubly robust methods. The simulation studies show that the proposed method improves the performance of the previous doubly robust method.

In my second project, I construct the hypothesis test to compare the treatment effects for multiple treatment groups in observational studies. Comparable to the randomized controlled trials, I proposed the weighted χ^2 test for categorical outcome variables and weighted F test for continuous outcome variables, in order to test the overall group difference. The weight for a subject is the inverse of the probability

for the subject receiving the assigned treatment given her/his own covariates. The simulation study shows that the proposed weighted tests could control for the family-wise error rate, while the traditional tests inflate the type I error rate.

In my third project, under the context of time dependent outcomes, I develop statistical methods to estimate ATE when there are a large number of potential confounding variables and censoring. Literature has suggested that the propensity score model should include the variables related to outcomes to obtain a more accurate ATE estimation. I propose to use variable selection method in outcome model to obtain the variables for propensity score estimation. I also incorporate the censoring data (informative or noninformative) via an inverse probability of uncensoring weighting in the ATE estimation. I propose the doubly weighting method coupled with the variable selection method to estimate ATE. I carry out the simulation studies to illustrate the advantages to use the proposed method.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Estimation of average treatment effects among multiple treatment groups by using an ensemble approach

In observational studies, generalized propensity score (GPS) based statistical methods, such as inverse probability weighting (IPW) and doubly robust (DR) method, have been proposed to estimate the average treatment effect (ATE) among multiple treatment groups. In this project, I investigate the GPS-based statistical methods to estimate treatment effects from two aspects. The first aspect of my investigation is to obtain an optimal GPS estimation method among four competing GPS estimation methods by using a rank aggregation approach. I further examine whether the optimal GPS based IPW and DR methods would improve the performance for estimating ATE. It is well known that the DR method is consistent if either the GPS or the outcome models are correctly specified. The second aspect of my investigation is to examine whether the DR method could be improved if I ensemble outcome models. To that end, bootstrap method and rank aggregation method are used to obtain the ensemble optimal outcome model from several competing outcome models, and the resulting outcome model is incorporated into the DR method, resulting in an ensemble DR method. Extensive simulation results indicate that the ensemble DR method provides the best performance in estimating the ATE regardless of the method used for estimating GPS. I illustrate the proposed methods using the MarketScan healthcare insurance claims database to examine the treatment effects among

three different bones and substitutes used for spinal fusion surgeries. The ensemble DR method coupled with the optimal GPS estimation method is recommended for ATE estimation.

1.2 Weighted χ^2 test and F test for multiple group comparisons in observational studies

Although χ^2 test and F test are commonly used for multiple group comparisons in experimental data, these methods cannot be directly used to examine group differences in observational studies because of the confounding factors. Since the seminal work by Rosenbaum and Rubin (1983), propensity-score-based inverse probability weighting (IPW) method has become one of the most popular methods for estimating ATE. However, the IPW method has only been applied to compare pairs among multiple treatment groups without controlling the family-wise error rate (FWER). In this project, I propose to examine whether there is an overall significant group difference using a weighted χ^2 test for a categorical outcome variable and a weighted F test for a continuous outcome variable. Only if there is an overall significant group difference, the pairs of interests are further examined. Alternatively, Bonferroni correction is applied to control the FWER for multiple group comparisons. Our extensive simulation studies show that the proposed methods can control the FWER, while the traditional tests have an inflated type I error rate. To illustrate the practical usage of the proposed tests, we apply the proposed weighted χ^2 test to investigate whether fruit/vegetable intakes are associated with heart attack using the 2015 Kentucky behavioral risk factor surveillance system dataset, and we apply the weighted F test to examine the effect of physical/recreational exercise on weight gain using the national health and nutrition examination survey I epidemiological follow-up study dataset.

1.3 Estimation of average treatment effect for time dependent outcomes

In observational studies, there are growing interests in estimating ATE with time dependent outcomes, such as time to event data, and the lifetime medical cost since diagnosis of a disease. Inverse probability of treatment weighting (IPTW) has been one of the most popular approaches to provide the consistent estimate of ATE, provided that the propensity score model is correctly specified. It is often a practice that researchers include all the measured covariates in the propensity score model, which inflates the variation of the ATE estimate if there are a large number of irrelevant variables. Thus, our first investigation is to use the lasso method to select the covariates for the propensity score model. In addition, the censoring, especially the informative censoring, needs to be accounted. I use the inverse probability of uncensoring weighting approach to account for the censoring, recovering to the situation that the entire study population are completely observed. To adjust confounding variables and censored observations, I propose the doubly weighting method to estimate ATE. The inverse of propensity score is estimated based on the selected covariates, and the inverse of probability of being uncensored is estimated from the Kaplan-Meiers estimator or Cox proportional hazard model. In the simulation study, I compare the performance of the doubly weighting method to the IPTW method. The simulation results show that the proposed technique performs well when there is a large number of covariates and there are censored observations. At last, we use the SEER-Medicare data to create a cohort of pancreas patients whose diagnose dates were between 2006 and 2013, and I apply the doubly weighting method to estimate the mean survival time under different treatment schemes.

CHAPTER 2

ESTIMATION OF AVERAGE TREATMENT EFFECTS AMONG MULTIPLE TREATMENT GROUPS BY USING AN ENSEMBLE APPROACH[1]

[1]

2.1 Introduction

Randomized controlled trials (RCT) are considered as the gold standard to determine the treatment effect between different treatment groups. In an RCT, the subjects are randomly assigned to different treatment groups and all confounding baseline covariates, either measured or unmeasured, are assumed to be balanced. Therefore, the treatment effect can be directly estimated by the difference of observed group means (Friedman et al., 2010). However, conducting an RCT is not always feasible because of ethics, cost, and patient preferences (Horwitz, 1987). But with the availability of the observed data in natural health care settings, estimating the treatment effect based on observational studies becomes more practical. Under an observational study, the treatment received by a subject is more likely determined by the subject's characteristics and the doctor's preference, and the covariates between different treatment groups may be unbalanced. Thus, the difference between two treatment groups is not

[1]The work has been published in *Statistic in Medicine*. The citation is "Yan, X., Abdia, Y., Datta, S., Kulasekera, K., Ugiliweneza, B., Boakye, M., andKong, M. (2019). Estimation of average treatment effects among multiple treatment groups by using an ensemble approach.Statistics in Medicine, 38:2828–2846."

only attributed to the treatment received, but also other variables, such as subject's age and other health conditions (Rubin, 2004).

To assess the treatment effect under an observational study, Rosenbaum and Rubin introduced the idea of propensity score (Rosenbaum and Rubin, 1983). The term propensity score refers to the probability of treatment assignment conditional on the observed baseline covariates. Propensity score is also known as a balancing score, that is, conditioning on the same value of the propensity score, the covariates in the treatment and control groups are similar. The logistic regression method is commonly used to estimate the propensity score. Recently the covariate balancing propensity score (CBPS) has been proposed to use the logistic regression method to model treatment assignment while optimizing the covariate balance, taking into account the two characteristics of propensity score (Imai and Ratkovic, 2014). However, the logistic regression model may lead to a biased estimator of the propensity score if the model is misspecified. To alleviate the possibility of model misspecificity, several non-parametric techniques, such as random forest and generalized boosted model (GBM), are proposed to estimate the 126 propensity score (McCaffrey et al., 2004; Lee et al., 2010). Once the propensity score becomes known, numerous propensity score based methods could be used to estimate the treatment effect. These methods include matching, regression with propensity score as a covariate, stratification, inverse probability weighting (IPW), and the doubly robust (DR) method (Lee et al., 2010; Rosenbaum and Rubin, 1985; Rosenbaum, 1987; Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004).

The propensity score framework was initially developed to assess treatment effect between two treatment groups (Rosenbaum and Rubin, 1983). Imbens extended the framework to estimate treatment effects among multiple treatment groups via the generalized propensity score (GPS) (Imbens, 2000). The GPS is defined as the conditional probability of receiving each treatment given pre-treatment variables (Im-

bens, 2000; Yang et al., 2016). Multinomial logistic regression is commonly used to estimate the GPS. The CBPS method could also be available for multiple treatment groups (Imai and Ratkovic, 2014). McCaffrey et al. proposed to use GBM to improve the estimate of the GPS and to balance the covariates (McCaffrey et al., 2013).

In this article, we investigate the GPS based statistical methods to estimate treatment effects from two aspects. Note that GPS could be estimated by parametric methods such as multinomial logistic regression and CBPS method, or non-parametric methods such as GBM and random forest. One important role of GPS is to balance the covariates. Thus, the first aspect of our investigation is to obtain an optimal GPS estimation method which is optimal in balancing covariates. We propose to use a rank aggregation approach to rank different GPS estimation methods based on their performance in balancing the covariates (Pihur et al., 2009). We further examine whether the optimal GPS based IPW and DR methods would improve the performance for estimating the average treatment effect (ATE), which is defined as the mean of individual causal effects for the whole population. It is well known that the DR method is consistent if either the GPS or the outcome models are correctly specified (Abdia et al., 2017). Hence, the second aspect of our investigation is to examine whether the ATE estimates based on the DR method could be further improved if we ensemble outcome models. To that end, when the outcome variable is continuous, we ensemble the commonly used outcome regression models, such as multiple linear regression model, random forest, and GBM, to form a better outcome model. Bootstrap method and rank aggregation method are used to obtain the ensemble optimal outcome model (Pihur et al., 2009; Datta et al., 2010). We further examine the performance of this ensemble DR method in estimating ATE.

Our paper is structured as follows. In Section 2.2, we present the assumptions of causal inference for multiple treatment groups and develop an optimal method to estimate the GPS. In Section 2.3, we first present the currently existing GPS based

methods for estimating ATE among multiple treatment groups. We then develop an ensemble DR (enDR) method by obtaining an ensemble adaptive optimal outcome model. In Section 2.4, extensive simulations are carried out to examine the performances of these proposed methods. In Section 2.5, a case study is presented to examine the treatment effects among three different bones and substitutes used for spinal fusion surgeries. Finally, Section 2.6 is devoted to a discussion.

2.2 Basic assumptions for causal inferences and GPS estimating methods

2.2.1 Notation and assumptions

Imbens (Imbens, 2000) outlines the framework for estimating the treatment effects via the generalized propensity score (GPS) when there are multiple treatment groups. Let X denote the vector of p pre-treatment covariates for a subject in the study, T and Y denote, respectively, the treatment received and the observed outcome for the subject. Suppose that there are M treatments to be considered. Each subject would have had M potential outcomes $(Y(1), Y(2), ..., Y(M))$, where $Y(t)$ would be the outcome if the subject receives the treatment t, $t \in \{1, \cdots, M\}$. However, each subject can only receive one treatment, say $T = t$, thus the observed outcome is the potential outcome corresponding to the treatment assigned, say $Y(t)$. Let (X_i, T_i, Y_i) denote the triplet of random variables for the i^{th} subject, where $i = 1, ..., n$. The GPS is the conditional probability of receiving a particular treatment given pre-treatment covariates (Imbens, 2000), which can be written as:

$$p(t|X) = \Pr(T = t|X), \text{ for } t = 1, \cdots, M. \tag{2.1}$$

To estimate treatment effects, the following assumptions are required (Yang et al., 2016; McCaffrey et al., 2013):

- Positivity (sufficient overlap): a subject has a non-zero probability of receiving each treatment. Mathematically, it is written as:

$$0 < \Pr(T = t|X) < 1, \text{ for } t = 1, ..., M \text{ and } X. \tag{2.2}$$

  Here $\sum_{t=1}^{M} \Pr(T = t|X) = 1$.

- Independence condition (i.e., weak unconfoundedness): assignment to treatment t is independent of the potential outcome $Y(t)$ given pre-treatment variables X, that is

$$D(t) \perp\!\!\!\perp Y(t)|X, \text{ for all } t, \tag{2.3}$$

  where $D(t)$ is 1 if $T = t$, and 0 otherwise. The independence condition implies that $D(t) \perp\!\!\!\perp Y(t)|p(t|X)$ for all t(Imbens, 2000; McCaffrey et al., 2013).

Under the assumptions of positivity and independence condition, it is possible to obtain an unbiased estimator of ATE by conditioning on the GPS rather than the entire covariate vector X. In the following, we first introduce four different GPS estimation methods: multinomial logistic regression, CBPS, random forest, and GBM. Then we develop an optimal method to estimate GPS determined by the performance in balancing covariates.

Four different GPS estimation methods

(i) *Estimating GPS using multinomial logistic regression*

Multinomial logistic regression is a commonly used method to estimate the

GPS. If we set $T = 1$ as the reference group, the multinomial logistic regression fits $M - 1$ regression equations:

$$\ln\left(\frac{p(t|X)}{p(1|X)}\right) = \ln\left(\frac{\Pr(T = t|X)}{\Pr(T = 1|X)}\right) = \beta_0^{(t)} + X'\beta_1^{(t)}, \text{ for } t = 2, \cdots, M. \tag{2.4}$$

Here $\sum_{t=1}^{M} p(t|X) = \sum_{t=1}^{M} \Pr(T = t|X) = 1$. The parameters $\beta_0^{(t)}$ and $\beta_1^{(t)}$ $(t = 2, \cdots, M)$ are estimated by maximizing the likelihood function based on observed covariates and treatment selection data. From Equation (2.4), we can get

$$\hat{p}(t|X) = \hat{p}(1|X)e^{\hat{\beta}_0^{(t)} + X'\hat{\beta}_1^{(t)}}, \text{ for } t = 2, \cdots, M. \tag{2.5}$$

Combining with the constraint $\sum_{t=1}^{M} \hat{p}(t|X) = 1$, the GPS can be obtained as follows:

$$\hat{p}(1|X) = \frac{1}{1 + \sum_{t'=2}^{M} e^{\hat{\beta}_0^{(t')} + X'\hat{\beta}_1^{(t')}}}, \tag{2.6}$$

and

$$\hat{p}(t|X) = \frac{e^{\hat{\beta}_0^{(t)} + X'\hat{\beta}_1^{(t)}}}{1 + \sum_{t'=2}^{M} e^{\hat{\beta}_0^{(t')} + X'\hat{\beta}_1^{(t')}}}, \text{ for } t = 2, \cdots, M. \tag{2.7}$$

(ii) *Estimating GPS using the covariate balancing propensity score (CBPS) method*

The propensity score plays two roles in an observational study: predict the probability of treatment assignment for each subject and balance the pre-treatment covariates between treatment and control groups. Imai and Ratkovic introduced the CBPS methodology to estimate the regression parameters in the treatment selection model using both score function and balance of covariates as estimating equations (Imai and Ratkovic, 2014). The treatment selection model in CBPS still uses a

logistic regression model. The CBPS methodology has been extended to multiple treatment groups and has been implemented in the "CBPS" package in R. In this article, we include CBPS as one candidate model in the optimal GPS estimation method developed in Section 2.2.3. We also combine CBPS with IPW and DR to estimate ATE.

(iii) *Estimating GPS using random forest*

Random forest is a tree-based method, which could capture complex interaction structures in the data. First, we generate B bootstrap samples from the original sample with replacement, then build a decision tree based on each bootstrap sample, resulting in B de-correlated trees (James et al., 2013; Friedman et al., 2001). During the buildup of these decision trees, when each split in a tree is conducted, a random sample of m covariates are chosen as split candidates from the full set of p covariates. m is usually taken as $\sqrt{p}$ for classification and $p/3$ for regression. The tree is grown until that the minimum node size is reached in each terminal node. A prediction at covariate $X = x$ is the class proportion (or the mean for regression) among training observations that fall into the same terminal node. The final class proportions are obtained from averaging those obtained from B trees. In our work, the GPS estimates based on random forest are calculated as the class proportions from these B trees, because treatment T is considered as a class factor. Since the performance of random forest may depend on tuning parameters m and B, we use a five-fold cross-validation method to select the optimal tuning parameters. In the simulation study, we choose m from the set $\{2, 4, 6, .., p-1\}$ and choose B from the set $\{1000, 5000\}$. The optimal tuning parameters could be obtained via the function *train()* in the "caret" package in R by setting *method="rf", metric="Accuracy"*. Given the selected tuning parameters, the *randomforest()* function in the "randomforest" package in R is utilized to estimate the GPS.

(iv) *Estimating GPS using generalized boosted model (GBM)*

The generalized boosted model (GBM) has been utilized to estimate the propensity scores for two groups (McCaffrey et al., 2004) and GPS for multiple treatment groups (McCaffrey et al., 2013). Different from the random forest, GBM does not involve bootstrap sampling. Instead, each tree is fit on the modified version of the original data. GBM starts from a simple classification (or regression) tree with d splits, that is, d+1 terminal nodes. GBM grows the trees sequentially: the new tree is chosen to provide the best fit to the residuals of the model from the previously grown trees. When adding the new tree, the contribution of each new tree is scaled by a factor less than one to improve the smoothness of the resulting model and the overall fit (McCaffrey et al., 2004, 2013; James et al., 2013; Friedman et al., 2001). GBM has three tuning parameters: the number of trees (or the iterations), the shrinkage parameter, and the number of splits in each tree. McCaffrey and his colleagues proposed to select optimal iterations that minimize absolute standardized mean differences for covariates (i.e., measure of covariate balance) (McCaffrey et al., 2004, 2013). This technique has been implemented in the function *mnps()* in the "twang" package in R for multiple treatment groups, which is used in this article to estimate the GPS. In the simulation study in Section 2.4 and the case study in Section 2.5, we set the arguments *n.trees=5000, stop.method="es.mean"* in the *mnps()* function.

2.2.2 An optimal GPS estimation method

Recall that GPS plays an important role in balancing covariates. The balance of a covariate is examined by the absolute standardized mean difference (ASMD) (McCaffrey et al., 2004). The ASMD statistic for t^{th} treatment group and j^{th} covariate

is given by

$$ASMD_j^{(t)} = \frac{|\bar{X}_{.j}^{(t)} - \bar{X}_{.j}|}{\hat{\sigma}_j}, \tag{2.8}$$

where $t = 1, \cdots, M$; $j = 1, \cdots, p$; and $\bar{X}_{.j}^{(t)} = \sum_{i=1}^{n} I_{\{T_i=t\}} w(t; X_i) X_{ij} / \sum_{i=1}^{n} I_{\{T_i=t\}}$ $w(t; X_i)$ is the weighted mean of the j^{th} covariate in the t^{th} group with $w(t; X_i)$ $= 1/p(t|X_i)$. Here $\bar{X}_{.j}$ and $\hat{\sigma}_j$ are the unweighted mean and standard deviation for the j^{th} covariate pooled across all treatment groups, respectively. It is noted that there are $M \times p$ ASMD scores for a study with M treatment groups and p covariates. We may summarize $M \times p$ ASMD scores to p ASMD scores by taking the average of $ASMD_j^{(t)}$s over M treatments for each covariate, that is, set $ASMD_j = \frac{1}{M} \sum_{t=1}^{M} ASMD_j^{(t)}$. Alternatively, we can summarize the $M \times p$ ASMD scores to p $ASMD$ scores by taking the maximum of $ASMD_j^{(t)}$ value over all treatment t $(t = 1, 2, ..., M)$, that is, $ASMD_j = \max_{1 \le t \le M} ASMD_j^{(t)}$ (McCaffrey et al., 2004). Thus, we get p ASMD scores associated with p covariates for each GPS estimation method. In general, ASMD greater than 0.10 indicates that the covariate is unbalanced. Although the four GPS estimation methods presented in section 2.2 are commonly used to estimate GPS, ASMD scores from one GPS estimation method are not consistently better than those from the other GPS estimation methods. However, we can rank the performance of these four GPS estimation methods for the balance of each covariate. The GPS estimation method with the smallest ASMD is ranked as the first, the GPS estimation method with the second smallest $ASMD$ as the second, and etc. For the j^{th} covariate, they are ranked according to their associated $ASMD_j'$s. Thus we have p such lists of ranks for the full set of p covariates. We apply the rank aggregation method (McCaffrey et al., 2004) to determine which GPS estimation method is optimal in balancing all covariates. Mathematically, the optimal method is defined as the method

ranked first for a list δ which minimizes the weighted rank aggregation quantity:

$$\Phi(\delta) = \sum_{j=1}^{p} \widetilde{w}_j dist(\delta, L_j). \tag{2.9}$$

Here δ is any valid ordered list of size K, which is the total number of GPS estimation methods (i.e., K=4). $dist()$ is the Spearman's footrule distance which measures the distance between δ and L_j $(j = 1, ..., p)$ (McCaffrey et al., 2013), where L_j is the ranks of $ASMD_j'$s obtained from the K GPS estimation methods, where the smallest $ASMD_j$ among the K values is ranked as 1. Also, $\widetilde{w}_j$ is an appropriate weight, which provides great flexibility in the rank aggregation. In the current work, we set the weights to be 1. The rank aggregation method can be carried out by the *BruteAggreg()* function in the "RankAggreg" package in R. In this article, we investigated the performance of the optimal GPS estimation method when $ASMD_j$ is defined as the mean of $ASMD_j^{(t)}$s over t as well as the maximum of $ASMD_j^{(t)}$s over t $(t = 1, 2, ..., M)$. We refer to the former as the optimal method based on MinMean criteria (say, $Opt_{MinMean}$), and the latter as the optimal method based on the MinMax criteria (say, Opt_{MinMax}).

2.3 GPS based statistical methods for estimating ATE

The average treatment effect (ATE) of treatment t' relative to t'' is the comparison of mean outcomes, when the entire population had been assigned to the treatment t' versus had been assigned to the treatment t'' (McCaffrey et al., 2013; Abdia et al., 2017). Mathematically, it can be written as:

$$ATE_{t',t''} = E(Y(t') - Y(t'')) = E(Y(t')) - E(Y(t'')) = \mu_{t'} - \mu_{t''}. \tag{2.10}$$

In the following, we first present the two commonly used GPS based methods for estimating ATE: inverse probability weighting (IPW) and doubly robust method (DR). Then we propose an ensemble doubly robust method (enDR).

2.3.1 Inverse probability weighting method for estimating ATE

The inverse probability weighting (IPW) method is often used for two group comparisons. McCaffrey et al. (McCaffrey et al., 2013) extended the IPW to estimate ATE when there are multiple treatment groups. The idea behind IPW is to construct the pseudo entire population for treatment t by weighting the subjects in the observed subgroup t. The mean outcome of the entire population under the t^{th} treatment, say μ_t, can be estimated as (McCaffrey et al., 2013)

$$\hat{\mu}_{t,IPW} = \frac{\sum_{i=1}^{n} I_{\{T_i=t\}} w(t;X_i) Y_i}{\sum_{i=1}^{n} I_{\{T_i=t\}} w(t;X_i)}, \tag{2.11}$$

where $w(t;X_i) = 1/p(t|X_i)$. The ATE between treatment t' and t'' can be estimated as difference between $\hat{\mu}_{t',IPW}$ and $\hat{\mu}_{t'',IPW}$, say $\Delta_{IPW}(t',t'') = \hat{\mu}_{t',IPW} - \hat{\mu}_{t'',IPW}$.

Suppose the GPS is known; then, the variance of $\hat{\mu}_{t',IPW} - \hat{\mu}_{t'',IPW}$ can be approximated as $n^{-2}\sum_{i=1}^{n} I_{i,IPW}^2$,(Lunceford and Davidian, 2004) where

$$I_{i,IPW} = \frac{I_{\{T_i=t'\}}(Y_i - \hat{\mu}_{t',IPW})}{p(t'|X_i)} - \frac{I_{\{T_i=t''\}}(Y_i - \hat{\mu}_{t'',IPW})}{p(t''|X_i)}. \tag{2.12}$$

As an alternative, the variance of $\Delta_{IPW}(t',t'')$ can be obtained by a bootstrap resampling method (Davison and Hinkley, 1997).

2.3.2 Doubly robust method for estimating ATE

The doubly robust (DR) estimator is an amendment to the IPW estimator (Robins et al., 1994). The DR estimator involves an outcome regression model for the outcome variable and a treatment selection model for estimating the GPS. DR estimator

remains consistent if either the GPS model or the outcome regression model is correctly specified. The DR estimator for the t^{th} treatment group is given by (Robins et al., 1994)

$$\hat{\mu}_{t,DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{I_{\{T_i=t\}}Y_i - (I_{\{T_i=t\}} - \hat{p}(t|X_i))m_t^{(DR)}(X_i)}{\hat{p}(t|X_i)}, \tag{2.13}$$

where $m_t^{(DR)}(X)$ is the outcome regression model for the outcome variable Y on X for the t^{th} treatment group, and $\hat{p}(t|X_i)$ is the estimated probability for the i^{th} subject to be assigned to t^{th} treatment group. The ATE between treatment t' and t'', $\Delta_{DR}(t', t'')$, can be estimated by the difference between $\hat{\mu}_{t',DR}$ and $\hat{\mu}_{t'',DR}$.

The variance of the estimated treatment effect $\Delta_{DR}(t', t'')$ can be estimated by $n^{-2}\sum_{i=1}^{i=n} I^2_{i,DR(t',t'')}$ (Lunceford and Davidian, 2004), where

$$\begin{aligned} I_{i,DR(t',t'')} = & \frac{I_{\{T_i=t'\}}Y_i - (I_{\{T_i=t'\}} - \hat{p}(t'|X_i))m_{t'}^{(DR)}(X_i)}{\hat{p}(t'|X_i)} \\ & - \frac{I_{\{T_i=t''\}}Y_i - (I_{\{T_i=t''\}} - \hat{p}(t''|X_i))m_{t''}^{(DR)}(X_i)}{\hat{p}(t''|X_i)} \\ & - \Delta_{DR}(t', t''). \end{aligned} \tag{2.14}$$

Again, an alternative method for estimating the variance would be a bootstrap resampling method.

2.3.3 Ensemble doubly robust method for estimating ATE

It is known that a doubly robust (DR) estimator is valid when either the GPS model or the outcome regression model is specified correctly. In the DR estimator, a multiple linear regression model is often used to model the outcome variable. However, a true outcome model is generally unknown and a multiple linear regression model may be restrictive in its model form and how variables enter into the model, where interaction and higher order terms are usually not included. As alternatives, machine

learning methods, such as GBM and random forest, can incorporate the higher order and interaction terms of the confounding variables. These methods may have the potential to provide better predictions for the outcome variable than multiple linear regression model. However, multiple linear regression model should not be excluded from candidate outcome models. Instead, we consider all potential outcome models, such as multiple linear regression model, random forest and GBM. We obtain a predicted outcome, according to an aggregation of many bootstrap samples over these potential outcome models. The idea can be traced back to the adaptive optimal ensemble method via rank aggregation (Datta et al., 2010; Shah et al., 2014). We adopt this concept to obtain an optimal outcome model, then further incorporate it into the doubly robust method to estimate the treatment effects. We call this method as the ensemble doubly robust (enDR) method. It is well known that the performance of GBM and random forest depends on their tuning parameters. The tuning parameters for these methods are selected based on a five-fold cross-validation method via the *train()* function in the "caret" package in R based on the original observed sample. We then use the same tuning parameters in the B bootstrap samples in the following algorithm for enDR:

1. Obtain the b^{th} bootstrap sample from the original observed sample. The bootstrap sample are divided into M subgroups based on the treatment assignment, denoted by $G_1^{(b)}, ..., G_M^{(b)}$. The out-of-bag (OOB) sample are also divided into M subgroups denoted by $G_1^{(OOB)}, ..., G_M^{(OOB)}$. For a specific treatment group, say the t^{th} group, K different outcome estimation models (e.g., K=3 for multiple linear regression, random forest, and GBM) are constructed based on the sample $G_t^{(b)}$. The performances of these K methods for predicting the outcome of t^{th} treatment group are ranked based on their prediction errors for $G_t^{(OOB)}$, resulting in an ordered list L_t, where the method with the smallest prediction error among the K values is ranked as 1. By examining their prediction errors

across all treatments, we form M ordered lists of size K, say $L_1, L_2, ..., L_M$.

2. The M ordered lists in Step 1 are aggregated using the weighted rank aggregation method, which minimizes the weighted rank aggregation quantity: $\Phi_{DR}(\delta) = \sum_{t=1}^{M} \widetilde{w}_t dist(\delta, L_t)$. The model at the top of the resulting list is considered as the best model. The overall rank is obtained by using the function *BruteAggreg()* in the "RankAggreg" package in R (Datta et al., 2010; Shah et al., 2014). We predict the M potential outcomes for each subject in the original sample using the best model selected in the bootstrap sample.

3. Repeat Step 1 to Step 2 B times (say, $B = 100$) and average these B sets of predicted outcome values to get the ensemble outcome estimate $m_t^{(enDR)}(X_i), i = 1, ..., n$, and $t = 1, ..., M$. Once we obtain the estimates of potential outcome $m_t^{(enDR)}(X_i)$, we replace $m_t^{(DR)}(X_i)$ with $m_t^{(enDR)}(X_i)$ in Equation (2.13) to get the ensemble DR estimate for the t^{th} treatment group:

$$\hat{\mu}_{t,enDR} = \frac{1}{n} \sum_{i=1}^{n} \frac{I_{\{T_i=t\}} Y_i - (I_{\{T_i=t\}} - \hat{p}(t|X_i)) m_t^{(enDR)}(X_i)}{\hat{p}(t|X_i)}, \tag{2.15}$$

The ATE between treatment t' and t'' (say, $\Delta_{enDR}(t', t'')$) can be estimated by the difference between $\hat{\mu}_{t',enDR}$ and $\hat{\mu}_{t'',enDR}$. The variance estimator of $\Delta_{enDR}(t', t'')$ can be obtained using the same estimator as for $\Delta_{DR}(t', t'')$ but using $m_t^{(enDR)}(X_i)$ instead of $m_t^{(DR)}(X_i)$ in Equation (2.14). Alternatively, the variance of $\Delta_{enDR}(t', t'')$ can be estimated by a bootstrap resampling method.

Remark 1: Random forest and GBM in the ensemble outcome model in Step 1. Random forest and GBM used in the ensemble outcome model follow the same algorithm as described in Section 2.2 for GPS model. However, random forest and GBM in the GPS model are for classification, and the class proportions among the observations that fall into a terminal mode are the estimated GPS. Random forest and GBM in the outcome model are for regression, and the mean of the observed

outcomes in a terminal node is the estimated outcome value (James et al., 2013; Friedman et al., 2001). We have used *mnps()* function in the "twang" package in R to estimate GPS and used *gbm()* function in the "gbm" package in R to obtain the outcome model.

Remark 2: Bootstrap variance estimator. When the GPS estimation method is coupled with the ensemble outcome model, the variance estimator in Equation (2.14) may not capture the variability due to the estimation of GPS. Thus, the variance based on Equation (2.14) may underestimate the variance. As an alternative approach, the bootstrap resampling method may provide a more accurate variance estimate for ATE. To obtain a bootstrap variance estimator, we draw B^* bootstrap samples from the original observed sample. For each bootstrap sample, we calculate the GPS and repeat Step 1-Step 3 to obtain B^* ATE estimates. The bootstrap variance estimator is the variance of these B^* ATE estimates (Davison and Hinkley, 1997). The bootstrap variance theoretically captures the variability from estimating GPS as well as from estimating the potential outcomes.

Remark 3: Ensemble outcome model. In the literature,the g-computation method has been used to estimate the ATE (Austin, 2012), that is, the outcome model under each treatment is obtained, and the potential outcome for i^{th} subject is predicted, say $(\hat{Y}_i(1), \hat{Y}_i(2), ..., \hat{Y}_i(M))$. The $ATE_{tt'}$ estimator is simply $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i(t) - \hat{Y}_i(t'))$, and the variance can be estimated as $\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{Y}_i(t) - \hat{Y}_i(t') - ATE_{tt'})^2$ (Austin, 2012). The resulting ATE estimates based on only the ensemble outcome model (say, enOM) are reported in the subsequent sections for simulations and case study. The variance of the enOM-based ATE estimate can be obtained more accurately via bootstrap resampling method.

2.4 Simulation study

2.4.1 Simulation settings

Simulations are conducted to examine the performance of different ATE estimation methods. The simulation structures are similar to the ones reported in the literature (Lee et al., 2010; Setoguchi et al., 2008; Setodji et al., 2017). We generated 15 multivariate normal variables denoted by X=($X_{.1}, X_{.2}, ..., X_{.15}$) with mean zero, unit variance, and correlation structure satisfying corr($X_{.1}, X_{.5}$)=corr($X_{.3}, X_{.8}$)=corr($X_{.11}, X_{.13}$) = 0.2, and corr($X_{.2}, X_{.6}$)=corr($X_{.4}$, $X_{.9}$)=corr($X_{.12}, X_{.14}$) = 0.9. Nine of these covariates ($X_{.1}, X_{.3}, X_{.5}, X_{.6}, X_{.8}, X_{.9}, X_{.13}, X_{.14}, X_{.15}$) are dichotomized by assigning -0.5 to negative numbers and 0.5 to positive numbers. Without using extra notation, we still denote $X = (X_{.1}, X_{.2}, ..., X_{.15})$ as the resulting 15 covariates. Among these fifteen covariates, $X_{.1}$, $X_{.2}$, $X_{.3}$ and $X_{.4}$ are true confounding variables which are related to both the treatment variable and the outcome variable; $X_{.5}$, $X_{.6}$ and $X_{.7}$ are exposure variables which are only related to the treatment variable but not to the outcome variable; and $X_{.8}, X_{.9}$ and $X_{.10}$ are predictor variables which are only related to the outcome variable but not to the treatment variable. The remaining covariates from $X_{.11}$ to $X_{.15}$ are distractors which are related to neither treatment variable nor outcome variable. A digraph describing the relationship between the covariates and the responses are shown in Figure A1.4 in the Appendix as well as in the article by Lee et al (Lee et al., 2010).

To examine the performance of different ATE estimation methods, we constructed two sets of treatment selection models, which follow multinomial logistic regression models (2.16)-(2.18) in Table 2.1 but with different complexity between the variables X and the treatment assignment variable T. Let $M = 3$, and $T \in \{1, 2, 3\}$. The model between variables X and the treatment assignment T is also called the generalized propensity score (GPS) model. The setting for GPS_A stipulates that GPS

is a function of the linear combination of X. The setting for GPS_B stipulates that GPS has a complex function form of X, which includes higher order terms and interaction terms of the variables related to treatment selection. Similarly, we constructed two sets of outcome models shown in Table 2.1. Out_A stipulates that the outcome Y is associated with the variables X in a linear fashion, while Out_B stipulates that the outcome is associated with the variables X in a complex fashion, including higher order and interaction terms. By combining these GPS and outcome models, there are four simulation scenarios: AA (GPS_A and Out_A), AB (GPS_A and Out_B), BA (GPS_B and Out_A), and BB (GPS_B and Out_B). In the underlying outcome model, (τ_1, τ_2) are the parameters to capture the treatment effect. Given (τ_1, τ_2), for each GPS model and outcome model, we generated 1000 data sets of size n (e.g. n =1000) and used those to estimate treatment effects with the following steps:

1. Generate n realizations of $X = (X_{.1}, X_{.2}, \cdots, X_{.15})'$.

2. Calculate the treatment selection probabilities based on the underlying GPS model (i.e., GPS_A or GPS_B) in Table 2.1 and the realization of X generated in Step 1.

3. Generate n realizations of the treatment assignment T from the multinomial distribution using the treatment selection probabilities calculated in Step 2.

4. Generate n realizations of the outcomes Y based on the n realizations of X and T in Steps 1-3 and the underlying outcome model (i.e., Out_A or Out_B) in Table 2.4.1.

5. Given T and X, use the multinomial logistic regression, random forest, GBM and CBPS methods to estimate GPS, denoted by $GPS^{(Mul)}$, $GPS^{(RF)}$, $GPS^{(GBM)}$ and $GPS^{(CBPS)}$, respectively. Then, select the optimal GPS estimation methods based on the MinMean criteria and MinMax criteria described in Section

2.2.3. The GPS estimates corresponding to these two selected methods are denoted as $GPS^{(MinMean)}$ and $GPS^{(MinMax)}$. Thus, there are six sets of GPS estimates in total.

6. Given Y, T and X, construct the multiple linear regression model to predict the potential outcomes ($m_1^{(DR)}(X_i)$, $m_2^{(DR)}(X_i)$, $m_3^{(DR)}(X_i)$) used in Equation (2.13) for subject i ($i = 1, 2, ..., n$). To estimate the ensemble outcome, apply the algorithm described in Section 2.3.3, which includes multiple linear regression, random forest and GBM as candidate outcome models.

7. Estimate the ATEs and their standard errors using the 19 ATE estimation methods described in Section 2.3: six GPS-based IPW methods, six GPS-based DR methods, six GPS-based enDR methods, and one enOM. The outcome for DR is calculated based on multiple linear regression model, and the outcome for enDR is calculated based on the ensemble outcome model in Step 6.

8. Repeat Step 1 to Step 7 1000 times. Keep the ATE estimates and the standard error estimates for each ATE estimation method for each simulation run.

It should be noted that the multinomial logistic regression model used in Step 5 included all X in the model but in an additive form. Thus, when the true treatment assignments were generated from model GPS_A, the multinomial logistic regression model was correctly specified, although more variables were included in the model. However, when the true treatment assignments were generated from model GPS_B, the multinomial logistic regression model which only included the variables in an additive form, was a misspecified model. Similarly, in Step 6, the potential outcome $m_t^{(DR)}(X_i)$ used in Equation (2.13) for DR was predicted by a multiple linear regression model, which included all X and two treatment indicator variables in an additive fashion. Thus, the outcome regression model used in the DR was correctly specified when the true outcome variable was generated under model Out_A but was misspecified when

the true outcome variable was generated under model Out_B. For enDR, when the true outcome variable was generated under model Out_A, the proposed enDR method included the correctly specified outcome model (i.e., multiple linear regression model) in the candidate models to predict the potential outcome; however, when the true outcome variable was generated under model Out_B, the enDR method didn't include the correctly specified outcome model in the candidate outcome models.

Table 2.1: Four simulation scenarios with data generated under two different treatment selection models (i.e., GPS_A and GPS_B) and two outcome models (i.e., Out_A and Out_B).

Treatment selection model	
	$T \sim (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$
	$\Pr(T=1\|\widetilde{X}) = \frac{1}{1+\exp(\widetilde{X}'\beta^{(1)})+\exp(\widetilde{X}'\beta^{(2)})}$ (2.16)
	$\Pr(T=2\|\widetilde{X}) = \frac{\exp(\widetilde{X}'\beta^{(1)})}{1+\exp(\widetilde{X}'\beta^{(1)})+\exp(\widetilde{X}'\beta^{(2)})}$ (2.17)
	$\Pr(T=3\|\widetilde{X}) = \frac{\exp(\widetilde{X}'\beta^{(2)})}{1+\exp(\widetilde{X}'\beta^{(1)})+\exp(\widetilde{X}'\beta^{(2)})}$ (2.18)
GPS_A	$\widetilde{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)'$ $\beta^{(1)}$ = (0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7)' $\beta^{(2)}$ = (0.7, -0.35, 0.5, -0.5, -0.85, 0.35, 0.8)'
GPS_B	$\widetilde{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_2^2, X_4^2, X_7^2, X_1X_3, X_2X_4, X_3X_5, X_4X_6, X_5X_7, X_1X_6, X_2X_3, X_3X_4, X_4X_5, X_5X_6)'$ $\beta^{(1)}$ = (0.8, -0.25, 0.6, -0.4, -0.8, -0.5,0.7, -0.25, -0.4, 0.7, 0.4, -0.175, 0.3, -0.28, -0.4, 0.4, -0.175, 0.3, -0.2, -0.4) $\beta^{(2)}$ = (0.7, -0.35, 0.5, -0.5, -0.85, 0.35, 0.8, -0.35, -0.5, 0.8, 0.35, -0.245, 0.25, -0.35, -0.425, 0.35, -0.245, 0.25, -0.25, -0.425)
Outcome model	
	$Y \sim (X_1, X_2, X_3, X_4, X_8, X_9, X_{10})$ $Y = \widetilde{X}'\alpha + \tau_1 I_{\{T=2\}} + \tau_2 I_{\{T=3\}} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ (19)
Out_A	$\widetilde{X} = (1, X_1, X_2, X_3, X_4, X_8, X_9, X_{10})'$ α=(-3.85, 0.3, -0.36, -0.73, -0.2, 0.71, -0.19, 0.26)' (τ_1, τ_2)=(0, 0) or (τ_1, τ_2)=(0, 0.5)
Out_B	$\widetilde{X} = (1, X_1, X_2, X_3, X_4, X_8, X_9, X_{10}, X_2^2, X_4^2, X_{10}^2, X_1X_3, X_2X_4, X_3X_8, X_4X_9, X_8X_{10}, X_1X_9, X_2X_3, X_3X_4, X_4X_8, X_8X_9)'$ α=(-3.850, 0.300, -0.360,-0.730, -0.200, 0.710, -0.190, 0.260, 0.300, -0.730, -0.190, -1.925, 0.210, -0.180, -0.511,0.100, -1.925, 0.210, -0.180, -0.365, -0.100)' (τ_1, τ_2)=(0, 0) or (τ_1, τ_2)=(0, 0.5)
Four simulation scenarios	
	AA: Data generated from GPS_A and outcome model Out_A AB: Data generated from GPS_A and outcome model Out_B BA: Data generated from GPS_B and outcome model Out_A BB: Data generated from GPS_B and outcome model Out_B

2.4.2 Simulation results

For each simulation scenario, we generated data under two specifications of (τ_1, τ_2): (0, 0) and (0, 0.5). Under the underlying outcome regression model (2.19) in Table 2.1, the true ATE for group 2 versus 1 is τ_1, the true ATE for group 3 versus 1 is τ_2, and the true ATE for group 3 versus 2 is $\tau_2 - \tau_1$. The simulation results, in terms of boxplots of the 1000 estimated ATEs for each ATE estimation method, are presented in Figure 2.1 and Figure 2.2, for (τ_1, τ_2) equal to (0, 0) and (0, 0.5), respectively. There are 19 ATE estimation methods: six GPS-based IPW methods, six GPS-based DR methods, six GPS-based enDR methods, and one ensemble outcome model (enOM), which are presented in the x-axis in Figures 2.1-2.2. From Figures 2.1-2.2,

we conclude that (i) when the candidate GPS models included the correctly specified model, all 19 ATE estimation methods provided unbiased estimators (Scenarios AA and AB). However, the variabilities of IPW were larger than those of DR, enDR and enOM under Scenario AA (i.e., the model to predict the outcome was correctly specified), and the variabilities of IPW and DR were larger than those of enDR and enOM under Scenario AB (i.e., the model to predict the outcome was mis-specified); (ii) when the candidate GPS models didn't include a correctly specified GPS model, but the candidate outcome models did include a correctly specified model (Scenario BA), IPW may result in a biased estimators for ATE. However, DR, enDR, and enOM did provide unbiased estimators; (iii) when neither GPS candidate models nor outcome models included correctly specified models (Scenario BB), enDR and enOM provided unbiased estimators, but IPW and DR may provide biased estimators; (iv) the variabilities of ATE estimates based on enDR and enOM were usually smaller than those based on IPW for all four simulation scenarios, and the variabilities of ATE estimates based on enDR and enOM were smaller than those based on DR when the outcome model was not correctly specified (Scenarios AB and BB); (v) when the candidate outcome models didn't include the correctly specified models (Scenarios AB and BB), the variabilities of ATE estimates based on all methods are larger than those when the candidate outcome models included the correctly specified model (Scenarios AA and BA). In summary, our simulation results clearly indicate that (i) enDR and enOM had better performance (i.e., less bias and smaller variability) than IPW in all simulation scenarios; (ii) enDR and enOM had comparable performance with DR when outcome model was completely specified (Scenarios AA and BA), but enDR and enOM had better performance than DR when outcome model was not correctly specified (Scenarios AB and BB). Thus, enDR and enOM are recommended for estimating ATE.

We used four existing GPS estimation methods (i.e., multinomial logistic re-

gression model, random forest, GBM, and CBPS) along with two proposed optimal GPS estimation methods (i.e., $Opt_{MinMean}$ and Opt_{MinMax}). It is worthwhile to examine their performance in balancing covariates and in estimating ATE. The boxplots of ASMD scores for the first 10 covariates are presented in Figure 2.3 for ASMD scores based on MinMax criteria, and in Figure A1.5 for ASMD scores based on MinMean criteria. From simulation results, we can see that (i) the performances of the six GPS estimation methods were quite similar in balancing covariates for the four simulation scenarios (Figures 2.3 and A1.5); (ii) when the GPS model was correctly specified (Scenarios AA and AB), CBPS has slightly larger bias in estimating ATEs (Figures A1.1-A1.2) but with smaller variability of ATE estimates (Figures 2.1 and 2.2); (iii) although the optimal GPS estimation methods are comparable to CBPS in Scenarios AA and AB in balancing covariates (Figures 2.3 and A1.5), the optimal GPS methods are less biased than CBPS in estimating ATE (Figures A1.1-A1.2); (iv) when GPS model was not correctly specified (Scenarios BA and BB), the performance of the optimal GPS was better than multinomial logistic regression and CBPS in estimating ATE. Based on our simulation results, we conclude that the enDR coupled with the optimal GPS (i.e., Opt_{MinMax}) performs robust in estimating ATE regardless of the simulation scenarios.

To examine the performance of variance estimators proposed in Equation (2.12) for IPW method and Equation (2.14) for DR and enDR method, we estimated the standard error (SE) (i.e., the squared root of variance) for each ATE estimate for each sample. We summarized the 1000 estimated SEs by their mean, which are reported in Table 2.2 under column "SE" for Scenario BB, as well as in Figures A1.1-A1.3 for Scenarios AA, AB, and BA, respectively. In addition, we reported the mean of 1000 estimated ATEs (see the column "EST" in Tables 2.2 and A1.1-A1.3), which would be close to the true ATE if the ATE estimator was unbiased. We also reported the standard deviation of the 1000 estimated ATE (see the column "Emp.SE" in Ta-

bles 2.2 and A1.1-A1.3). By comparing the mean of 1000 estimated standard errors (SE) with the empirical standard deviation (Emp.SE), we can gauge the accuracy of the estimated standard errors. From Table 2.2 as well as A1.1-A1.3, we can see that the mean of SE estimates (see the column "SE") are almost always larger than the empirical standard deviation (see the column"Emp.SE") for IPW-type estimators, indicating that variance estimators are overestimated. The SEs for DR are close to Emp.SE for Scenarios AA, AB and BA, but not for Scenario BB. The SEs for enDR are close to Emp.SE for Scenarios AA and BA, but not for Scenarios AB and BB. The SEs for enOM seem not consistent with Emp.SE. To remedy this shortcoming, we applied the bootstrap resampling method to estimate the variance for Scenario BB. The results are presented in Table 2.2 under the column "BS.SE". By comparing the bootstrap standard error (see the column"BS.SE") with the empirical standard error (see the column "Emp.SE"), the bootstrap variance estimator is close to the empirical variance. Thus, the bootstrap variance estimator is more accurate, although the underlying computation is intensive.

Even though enDR and enOM have comparable or better performance than IPW and DR in all simulation scenarios, the bias for enDR and enOM for Scenario BB is still relatively large (e.g., 0.525 versus the true 0.5). For Scenario BB, we carried out the simulation with sample size 5000 (Table 2.3). From the simulation results, the estimates from enDR and enOM are close to the true values, while the estimates from IPW and DR have not been improved, particularly when GPS was estimated by multinomial logistic regression and CBPS in Scenario BB. In all simulation scenarios, enDR coupled with the optimal GPS estimation method performed slightly better than enOM, and therefore enDR coupled with the optimal GPS estimation method (i.e., Opt_{MinMax}) is recommended.

2.5 A case study

The MarketScan Commercial Claims and Encounters (MarketScan) Database contains de-identified, person-specific health data of reimbursed healthcare claims for employees, retirees, and their dependents of over 250 millions of employers and health plans. Our study team has purchased a custom MarketScan database related to neurological/neurosurgical conditions, which contains the insurance claims made by Medicare, Medicaid, and commercial insurance companies. Data used for this project covers data from years 2001 to 2011. We are interested in comparing the outcomes for three different bones and substitute used for spinal fusion surgeries: bone morphogenetic proteins (BMP), autograft, and allograft (Giannoudis et al., 2005; Gibson et al., 2002). BMP is a naturally occurring protein within our bodies which stimulates bone to form. During a fusion surgery, the spine surgeon places BMP on a sponge at the surgical site to cause the adjacent bones to fuse together. Before BMP, the traditional gold standard for bone graft material was an autograft, the patient's own hip bone. Limitations, however, exist regarding donor site morbidity and graft availability. Allograft (using bones harvested by a tissue bank) has been the most frequently chosen bone substitute and is regarded as the surgeon's second option. However, allograft possesses the risk of disease transmission. In this case study, we are particularly interested in examining the overall health care cost after the procedures in outpatient services. We consider to adjust the following confounding factors: (i) fusion type: inter-body fusion, posterior fusion and circumferential fusion; (ii) sex; (iv) age; (v) geographic region; (vi) types of insurance (i.e., Medicare, Medicaid, and commercial); and (vi) the Charlson comorbidity score.

We included 49,582 subjects in the study. Each subject had a spinal degenerative disease and was treated with only one of the three fusion bone materials: BMP, autograft, or allograft. Among these 49,582 subjects, 28,759 were female and 20,823

were male. For these 49,582 subjects, 6,135 insurance claims were from Medicare, 4,444 insurance claims were from Medicaid, and all others were from commercial insurance companies. Among all subjects, 9,599 were treated with BMP, 22,842 were treated with allograft and 17,141 were treated with autograft. We applied different statistical methods to compare the cost of the three groups with adjustment of the confounding factors. The results for group comparisons are reported in Table 2.4, and the balance of covariates are reported in Figure 2.4 for Opt_{MinMax} method and Figure A1.6 for $Opt_{MinMean}$ method.

In the case study, the ASMDs (i.e., the balance of the covariates) based on multinomial logistic regression, GBM and CBPS are below 0.1 (below the horizontal line in Figure 2.4), indicating that the covariates after adjustment were similar among the three groups. The optimal GPS based on MinMean criteria selected the multinomial logistic regression model while the optimal GPS based on MinMax criteria selected the GBM. The balance of covariates from the random forest is poor (Figures 2.4 and A1.5), even though we have selected the tuning parameter for the number of covariates among $\{2, 4, 6\}$ and tree size among $\{1000, 5000\}$ with the five-fold cross-validation method. The results for group comparisons are presented in Table 2.4 as ATE estimate and its estimated standard error for each ATE estimation method. The bootstrap resampling technique has been applied to obtain the standard errors (SE) for each ATE estimate. From Table 2.4, the ATE estimates based on the random forest are quite different from all other methods, which may be due to the unbalanced covariates. Based on the simulation studies in Section 2.4, when the optimal GPS estimation method was combined with the enDR method, the ATE is generally unbiased. We draw conclusions based on the ATE estimates from the enDR with the optimal GPS method (i.e., Opt_{MinMax}). Based on Table 2.4, the post-surgery outpatient cost for BMP is the highest compared to the allograft and autograft, since the cost for BMP is \$4110 (SE=\$1370) higher than autograft and \$3231 (SE=\$1279)

higher than allograft.

2.6 Discussion

In this article, we proposed to select optimal GPS estimation method in balancing covariates by using rank aggregation approach from the currently available GPS estimation methods, which include the multinomial logistic regression model, CBPS, random forest, and GBM. Based on the simulation results, the optimal GPS estimation method performs robust in estimating ATEs. Further, we also proposed the enDR method to improve the DR method with the idea of ensembling outcome models. To that end, bootstrap method and rank aggregation method are used to obtain the ensemble optimal outcome model from three possible models, and the resulting ensemble outcomes are incorporated into the DR method. As a byproduct, we also report the results based on g-computation method (Austin, 2012), which only uses the ensemble outcome models. Extensive simulation results indicate that the enDR method coupled with the optimal GPS estimation method (i.e., Opt_{MinMax}) provides the best performance in estimating ATE. We illustrate our methods using the MarketScan healthcare insurance claims database to examine the treatment effects among three different bones and substitutes used for spinal fusion surgeries.

Lunceford and Davidian (Lunceford and Davidian, 2004) developed variance estimator for ATE when GPS is known or estimated by parametric models. However, when the GPS was estimated by a machine learning method such as random forest, GBM, or the optimal GPS estimation method developed in this article, it is difficult to incorporate the variability from estimating GPS into the variance estimator for ATE. Instead, in the variance estimator (2.12) for IPW-type ATE estimator and (2.14) for the DR-type estimator, we ignored the variability from GPS estimation. The variance

estimators from (2.12) were similar to those obtained from the function *svyglm()* in the "survey" package in R (result not shown), which is the main package used in estimating the variance when machine learning method is used to estimate GPS (Lee et al., 2010; Lumley et al., 2004). To incorporate the variability from estimating GPS, the bootstrap resampling method is used to provide a more accurate variance estimator for ATEs.

The GPS plays dual roles in estimating ATE: modeling the probability that each subject is assigned to different treatment and balancing the covariates. Methods have been developed to estimate GPS with covariate balance built in as estimating equations in parametric models (e.g., "CBPS" package in R) or with covariate balance as an estimating criteria for selecting tuning parameters in the non-parametric approach (e.g., *mnps()* for GBM in the "twang" package in R). A GPS model with a better goodness-of-fit may not necessarily guarantee a better estimate for ATE. In the case study based on the MarketScan dataset, the random forest has a slightly larger correct classification rate (0.479) than the multinomial logistic regression model (0.474) and CBPS model (0.474). However, the random forest has the worst covariate balance (see Figures 2.4 and A1.6), resulting in highly biased estimators for ATEs. In our early investigation, we ensembled GPS estimation models by using rank aggregation and bootstrap methods (Abdia, 2016). However, the ensembled GPS method did not generate much improved ATE estimates. The ensemble outcome model developed in this article coupled with the optimal GPS method (i.e., Opt_{MinMax}) provided much improved estimates for ATEs. A reviewer has brought a double/debiased machine learning (ML) method to our attention (Chernozhukov et al., 2018). Double ML method uses the Neyman-orthogonal estimating equation and sample splitting strategy to estimate ATE for two groups (Chernozhukov et al., 2018). ML methods are used to estimate the nuisance relationship between X and T, as well as X and Y using one part of the data. Then the Neyman orthogonal estimating equation is used

to estimate ATE based on the remaining portion of the data. Double ML does not consider the balance of covariates. Double ML in estimating ATE for multiple groups may be worth investigating. However, it is beyond the scope of the current work.

Table 2.2: Simulation results for Scenario BB (i.e., GPS_B and Out_B), where EST, SE, and BS.SE are, respectively, the average of 1000 estimated ATEs, their estimated standard errors based on the formula (2.12) or (2.14), and their estimated standard errors based on bootstrap method. Emp.SE is the standard deviation of the 1000 estimated ATEs.

(τ_1, τ_2)=(0, 0)													
Comparison groups		2 vs 1				3 vs 1				3 vs 2			
True ATE		0				0				0			
		EST	Emp.SE	SE	BS.SE	EST	Emp.SE	SE	BS.SE	EST	Emp.SE	SE	BS.SE
IPW	Mul	0.196	0.223	0.249	0.209	0.248	0.230	0.242	0.209	0.052	0.117	0.175	0.117
	RF	-0.105	0.101	0.199	0.118	-0.004	0.106	0.192	0.114	0.101	0.101	0.189	0.110
	GBM	-0.107	0.106	0.196	0.104	-0.010	0.113	0.190	0.107	0.097	0.097	0.189	0.096
	CBPS	0.335	0.187	0.210	0.177	0.395	0.185	0.202	0.174	0.060	0.124	0.168	0.125
	MinMean	0.128	0.250	0.214	0.233	0.201	0.242	0.206	0.222	0.073	0.114	0.178	0.124
	MinMax	0.122	0.255	0.212	0.236	0.195	0.241	0.205	0.224	0.073	0.113	0.178	0.125
DR	Mul	0.488	0.283	0.255	0.246	0.545	0.283	0.249	0.242	0.058	0.119	0.111	0.115
	RF	0.175	0.115	0.133	0.121	0.203	0.109	0.126	0.116	0.027	0.090	0.119	0.087
	GBM	0.169	0.116	0.099	0.117	0.204	0.111	0.096	0.113	0.034	0.093	0.091	0.089
	CBPS	0.341	0.159	0.150	0.156	0.407	0.154	0.143	0.150	0.066	0.116	0.106	0.111
	MinMean	0.313	0.200	0.162	0.182	0.364	0.207	0.155	0.180	0.051	0.110	0.108	0.109
	MinMax	0.305	0.193	0.158	0.178	0.353	0.200	0.152	0.177	0.049	0.108	0.109	0.109
enDR	Mul	0.015	0.109	0.069	0.082	0.046	0.108	0.068	0.081	0.031	0.068	0.032	0.062
	RF	-0.014	0.071	0.040	0.070	0.017	0.068	0.039	0.069	0.031	0.065	0.034	0.060
	GBM	-0.019	0.072	0.034	0.071	0.015	0.069	0.034	0.069	0.034	0.066	0.028	0.060
	CBPS	-0.003	0.077	0.044	0.072	0.029	0.073	0.043	0.071	0.032	0.068	0.030	0.062
	MinMean	-0.004	0.079	0.047	0.074	0.028	0.076	0.046	0.072	0.032	0.067	0.032	0.062
	MinMax	-0.005	0.077	0.046	0.073	0.027	0.074	0.045	0.072	0.032	0.067	0.032	0.061
enOM		-0.022	0.076	0.023	0.074	0.017	0.072	0.024	0.070	0.039	0.069	0.015	0.035
(τ_1, τ_2)=(0, 0.5)													
True ATE		0				0.5				0.5			
		EST	Emp.SE	SE	BS.SE	EST	Emp.SE	SE	BS.SE	EST	Emp.SE	SE	BS.SE
IPW	Mul	0.199	0.210	0.246	0.204	0.745	0.212	0.240	0.204	0.546	0.116	0.175	0.117
	RF	-0.100	0.099	0.199	0.117	0.498	0.101	0.192	0.113	0.598	0.102	0.189	0.110
	GBM	-0.102	0.106	0.196	0.103	0.492	0.115	0.190	0.107	0.595	0.097	0.189	0.095
	CBPS	0.341	0.186	0.211	0.175	0.898	0.185	0.203	0.172	0.556	0.125	0.169	0.125
	MinMean	0.126	0.253	0.212	0.228	0.695	0.238	0.205	0.218	0.569	0.114	0.178	0.124
	MinMax	0.122	0.258	0.211	0.231	0.690	0.242	0.204	0.219	0.569	0.113	0.179	0.124
DR	Mul	0.497	0.249	0.251	0.243	1.047	0.247	0.246	0.238	0.550	0.116	0.111	0.115
	RF	0.179	0.111	0.133	0.121	0.702	0.107	0.126	0.116	0.523	0.089	0.119	0.087
	GBM	0.173	0.112	0.099	0.116	0.704	0.111	0.096	0.113	0.531	0.092	0.091	0.089
	CBPS	0.348	0.159	0.150	0.155	0.909	0.155	0.143	0.149	0.560	0.113	0.106	0.111
	MinMean	0.314	0.203	0.161	0.179	0.860	0.207	0.155	0.178	0.546	0.106	0.109	0.109
	MinMax	0.305	0.198	0.157	0.176	0.848	0.204	0.151	0.175	0.544	0.108	0.110	0.108
enDR	Mul	0.014	0.093	0.070	0.084	0.542	0.092	0.069	0.083	0.528	0.067	0.033	0.061
	RF	-0.015	0.070	0.040	0.069	0.514	0.069	0.040	0.068	0.529	0.063	0.035	0.059
	GBM	-0.019	0.071	0.034	0.070	0.512	0.070	0.034	0.068	0.532	0.063	0.029	0.060
	CBPS	-0.003	0.076	0.045	0.072	0.526	0.074	0.044	0.070	0.530	0.066	0.031	0.061
	MinMean	-0.004	0.077	0.047	0.073	0.525	0.076	0.047	0.071	0.529	0.065	0.032	0.061
	MinMax	-0.005	0.077	0.046	0.073	0.524	0.076	0.046	0.071	0.529	0.065	0.032	0.061
enOM		-0.022	0.076	0.023	0.073	0.515	0.074	0.024	0.072	0.537	0.067	0.015	0.063

Note: Mul indicates multinomial logistic regression model; RF indicates random forest; DR indicates doubly robust method; enDR indicates ensemble doubly robust method; enOM indicates ensemble outcome model.

Table 2.3: Simulation results for Scenario BB (i.e., GPS_B and Out_B) with sample size 5000, where EST and SE are, respectively, the average of 1000 estimated ATEs and their estimated standard errors. Emp.SE is the standard deviation of the 1000 estimated ATEs.

(τ_1, τ_2)		(0,0)									(0,0.5)								
Comparison groups		2 vs 1			3 vs 1			3 vs 2			2 vs 1			3 vs 1			3 vs 2		
True ATE		0			0			0			0			0.5			0.5		
		EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE
IPW	Mul	0.195	0.093	0.117	0.253	0.094	0.115	0.058	0.048	0.077	0.192	0.096	0.117	0.750	0.096	0.114	0.558	0.049	0.076
	RF	-0.038	0.042	0.087	0.020	0.042	0.085	0.058	0.039	0.086	-0.042	0.041	0.086	0.517	0.042	0.084	0.559	0.039	0.086
	GBM	-0.054	0.044	0.076	-0.006	0.047	0.076	0.049	0.038	0.075	-0.056	0.044	0.076	0.495	0.048	0.076	0.551	0.037	0.075
	CBPS	0.395	0.097	0.091	0.408	0.098	0.087	0.012	0.054	0.074	0.392	0.097	0.091	0.904	0.098	0.087	0.512	0.054	0.074
	MinMean	0.032	0.156	0.090	0.086	0.148	0.088	0.054	0.044	0.082	0.029	0.151	0.090	0.585	0.143	0.088	0.555	0.045	0.082
	MinMax	0.080	0.187	0.091	0.129	0.176	0.089	0.049	0.046	0.081	0.075	0.184	0.091	0.626	0.172	0.088	0.551	0.046	0.080
DR	Mul	0.500	0.110	0.119	0.560	0.109	0.117	0.060	0.049	0.049	0.494	0.112	0.119	1.055	0.111	0.116	0.560	0.049	0.049
	RF	0.143	0.041	0.060	0.162	0.040	0.058	0.019	0.034	0.056	0.137	0.041	0.060	0.658	0.041	0.058	0.521	0.033	0.056
	GBM	0.103	0.040	0.051	0.132	0.040	0.050	0.029	0.035	0.049	0.100	0.041	0.051	0.631	0.040	0.050	0.531	0.034	0.049
	CBPS	0.344	0.068	0.070	0.418	0.067	0.066	0.073	0.048	0.048	0.340	0.069	0.070	0.913	0.067	0.066	0.574	0.048	0.048
	MinMean	0.211	0.151	0.068	0.242	0.166	0.066	0.031	0.042	0.053	0.210	0.150	0.069	0.744	0.166	0.067	0.533	0.043	0.053
	MinMax	0.242	0.161	0.071	0.280	0.180	0.069	0.038	0.046	0.052	0.233	0.159	0.070	0.773	0.176	0.068	0.540	0.045	0.052
enDR	Mul	-0.003	0.019	0.015	0.006	0.020	0.015	0.009	0.017	0.007	-0.007	0.021	0.017	0.503	0.021	0.017	0.510	0.018	0.007
	RF	-0.006	0.015	0.008	0.003	0.016	0.008	0.009	0.017	0.007	-0.009	0.015	0.009	0.501	0.017	0.009	0.510	0.018	0.008
	GBM	-0.007	0.015	0.007	0.002	0.016	0.008	0.010	0.017	0.007	-0.011	0.015	0.008	0.500	0.017	0.008	0.510	0.017	0.007
	CBPS	-0.007	0.016	0.009	0.002	0.017	0.009	0.009	0.017	0.006	-0.011	0.016	0.010	0.499	0.018	0.010	0.510	0.018	0.007
	MinMean	-0.006	0.015	0.009	0.004	0.017	0.009	0.009	0.017	0.007	-0.009	0.016	0.010	0.501	0.017	0.010	0.510	0.018	0.008
	MinMax	-0.006	0.016	0.009	0.003	0.017	0.010	0.009	0.017	0.007	-0.009	0.016	0.010	0.501	0.017	0.010	0.510	0.018	0.007
enOM		-0.011	0.016	0.005	-0.001	0.017	0.006	0.010	0.017	0.004	-0.015	0.016	0.005	0.495	0.017	0.006	0.511	0.018	0.003

Note: Mul indicates multinomial logistic regression model; RF indicates random forest; DR indicates doubly robust method; enDR indicates ensemble doubly robust method; enOM indicates ensemble outcome model.

Table 2.4: ATE estimates and their standard errors for group comparisons based on the MarketScan data set with different GPS-based-methods (i.e., IPW, DR, enDR) and enOM. The GPS was estimated using multinomial logistic regression (Mul), random forest (RF), GBM, CBPS, the optimal GPS based on MinMean criteria, and the optimal GPS based on MinMax criteria, respectively. In each cell, the first number is the estimated ATE, and the second number is the estimated standard error based on the bootstrap method.

	IPW						DR						enDR						enOM
	Mul	RF	GBM	CBPS	MinMean	MinMax	Mul	RF	GBM	CBPS	MinMean	MinMax	Mul	RF	GBM	CBPS	MinMean	MinMax	
BMP vs Allograft	702	21806	867	797	702	867	666	-12014	873	663	666	873	795	-13867	879	801	795	879	912
	493	2235	489	496	491	495	491	35915	484	491	487	492	483	19874	484	483	482	485	483
BMP vs Autograft	4940	-3216	4094	5069	4940	4094	4957	-14204	4109	4955	4957	4109	4330	19393	4110	4340	4330	4110	4572
	1352	1475	1340	1348	1403	1348	1360	181489	1334	1360	1405	1350	1375	65673	1372	1378	1377	1370	1403
Allograft vs Autograft	4238	-25022	3227	4273	4238	3227	4291	-2189	3237	4291	4291	3237	3536	33260	3231	3539	3536	3231	3660
	1255	2536	1254	1255	1342	1267	1263	181655	1245	1263	1348	1274	1286	65242	1288	1288	1291	1279	1311

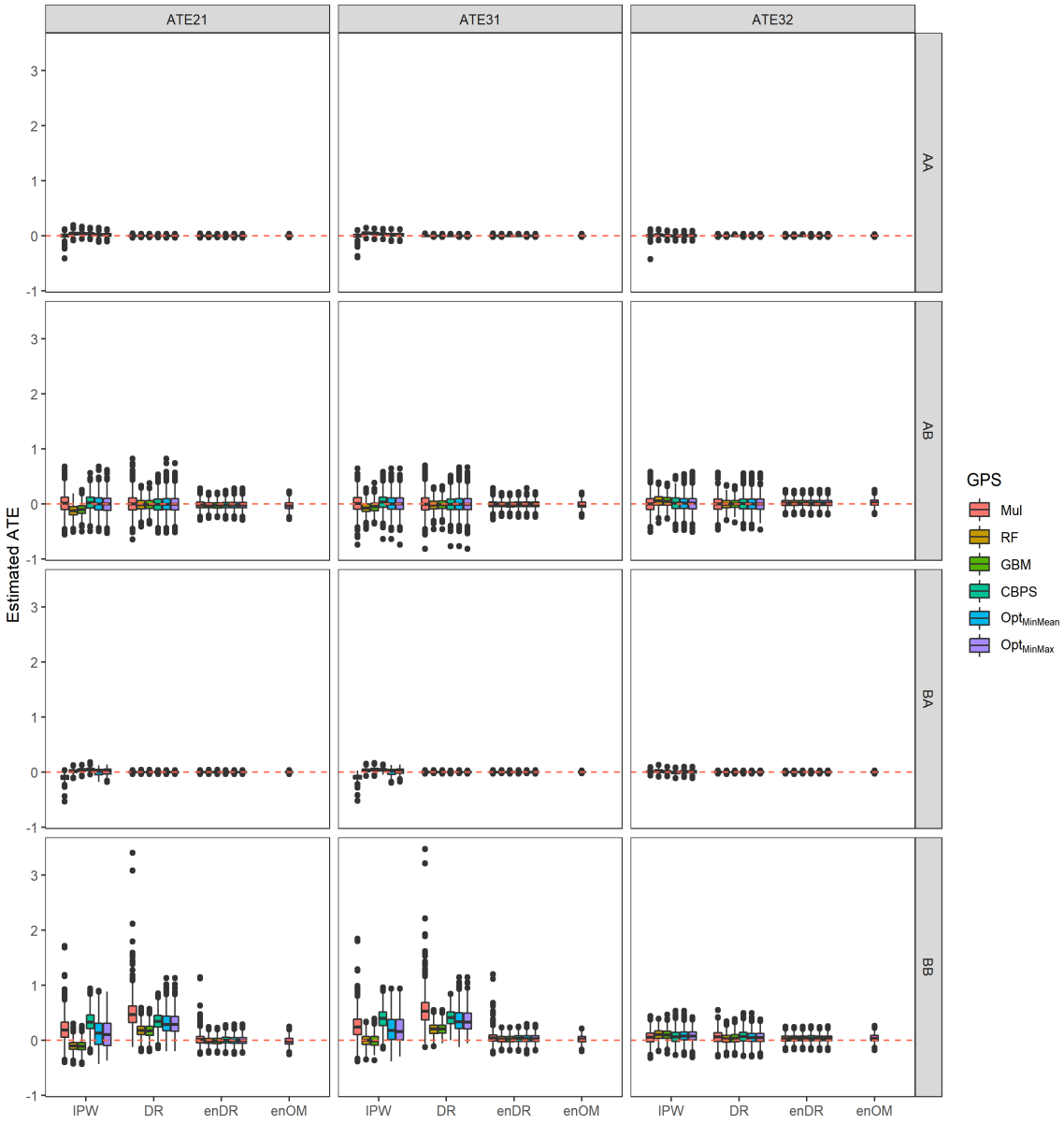

Figure 2.1: The boxplots of 1000 estimated ATEs for each of the 19 different methods (i.e., 6 IPWs, 6 DR, 6 enDR, and 1 enOM) under four different scenarios (i.e., AA, AB, BA and BB) with $(\tau_1, \tau_2) = (0, 0)$.

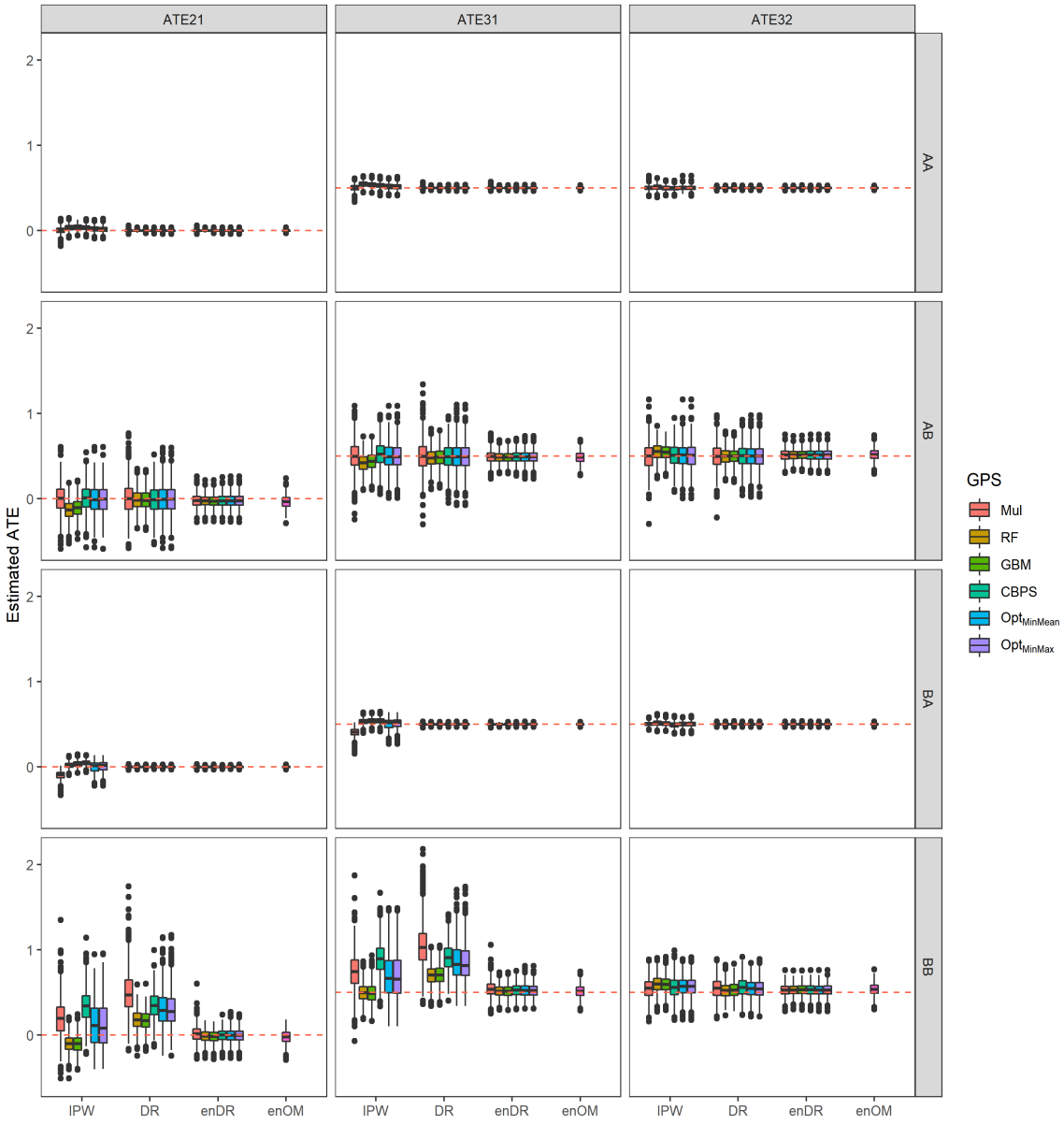

Figure 2.2: The boxplots of 1000 estimated ATEs for each of the 19 different ATE estimation methods (i.e., 6 IPWs, 6 DR, 6 enDR, and 1 enOM) under four different scenarios (i.e., AA, AB, BA and BB) with $(\tau_1, \tau_2) = (0, 0.5)$.

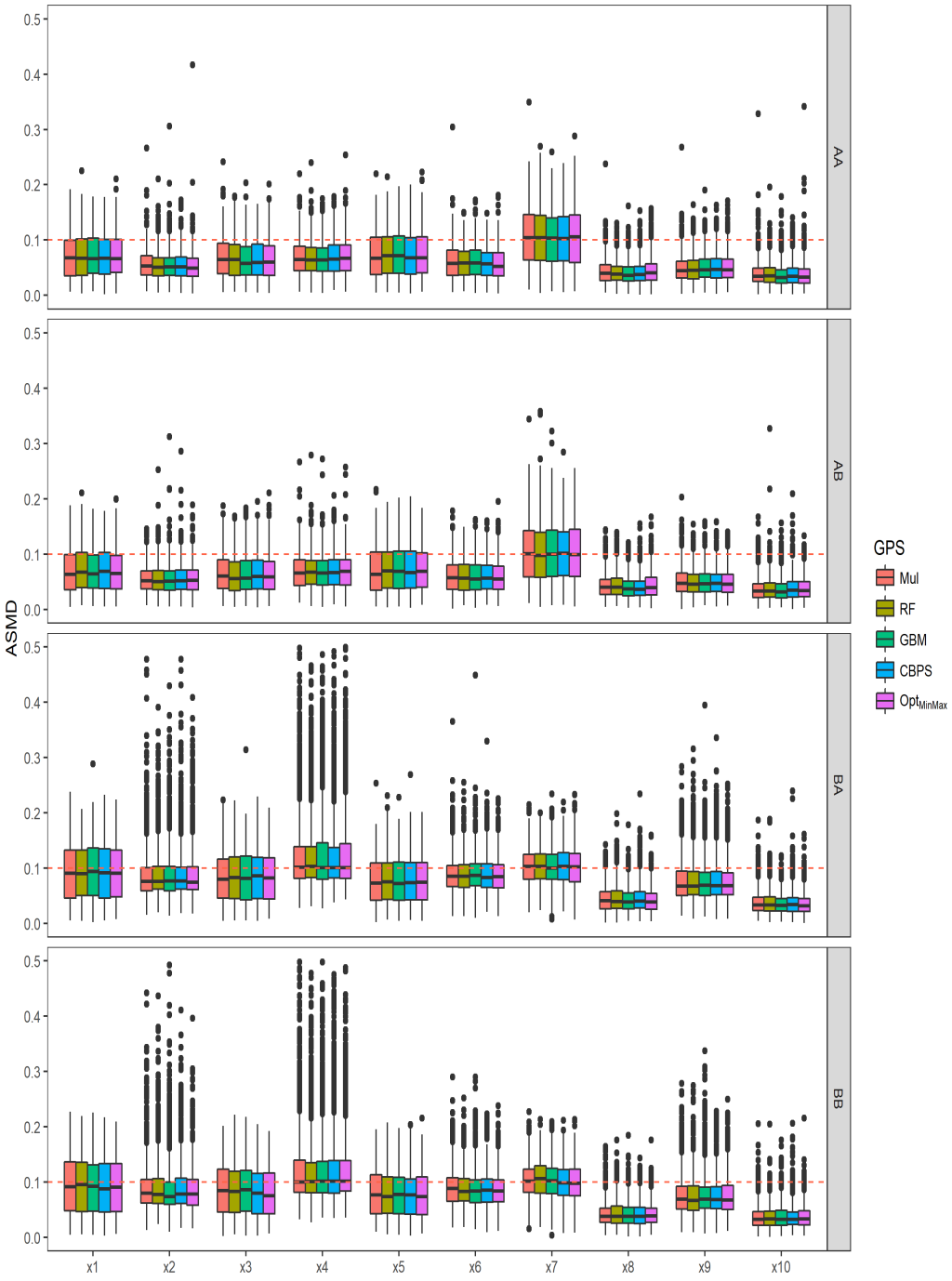

Figure 2.3: The boxplots of 1000 absolute standardized mean differences (ASMDs) based on MinMax criteria for four simulation scenarios under five different GPS estimation methods: multinomial logistic regression (Mul), random forest (RF), GBM, and the covariate balancing propensity score (CBPS), and the optimal GPS estimation method (Opt_{MinMax}), where a lower ASMD indicates a better balance of the covariates.

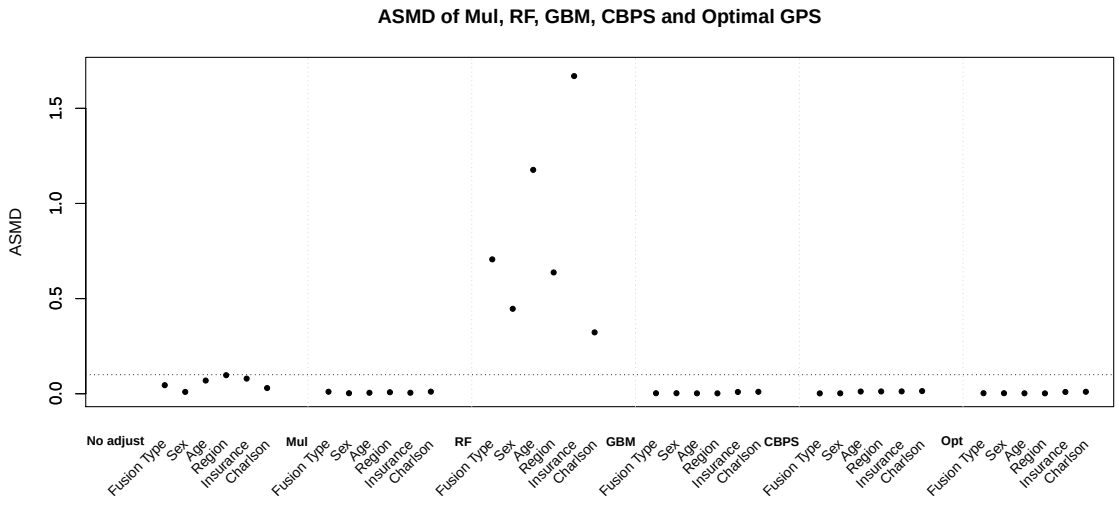

Figure 2.4: Absolute standardized mean differences (ASMDs) for the MarketScan dataset: ASMD without any adjustment (No adjust), ASMDs under four different GPS estimation methods (i.e., multinomial logistic regression (Mul), random forest (RF), GBM, CBPS) and the optimal GPS estimation method based on MinMax criteria (*Opt*). The covariates (fusion type, sex, age, region, insurance and Charlson comorbidity index) are included in the GPS and outcome model. The horizontal line for h=0.1 is the recommended cut-point on whether a covariate is balanced or not. A lower ASMD indicates a better balance of covariate.

CHAPTER 3

WEIGHTED χ^2 TEST AND F TEST FOR MULTIPLE GROUP COMPARISONS IN OBSERVATIONAL STUDIES

3.1 Introduction

There have been growing interests in comparing treatment effects among multiple treatment groups in biomedical studies (McCaffrey et al., 2013; Yan et al., 2019). In randomized controlled trials (RCT), subjects are randomly assigned to different treatment groups. Thus the subject's characteristics (i.e., covariates) are independent of his/her assigned treatment, that is, there is no confounding, which implies that the distribution of a covariate across different treatment groups are similar. One can directly estimate the mean outcome of each intervention by averaging the observed outcomes from subjects receiving the intervention (Horwitz, 1987; Rubin, 2004; Hernán and Robins, 2020). Thus, the difference of sample means provides a consistent estimate of the the average treatment effect (ATE) of the intervention comparing to the control. However, in an observational study, the treatment selection may be affected by the subject's characteristics. For example, in an observational study investigating the effect of the heart transplant on patients, patients with severe heart disease are more likely to undergo the heart transplant. Therefore, patients in the heart transplant group tend to have more severe heart disease conditions compared to those who do not have heart transplant (say no-heart-transplant group). As a result, comparing the average survival time between the heart transplant group and the no-heart-transplant group without considering the heterogeneity of heart disease

conditions can lead to erroneous conclusions. Thus, in observational studies, in order to correctly compare the treatment effects, one needs to adjust for the confounding factors that impact both the treatment selection and the outcome (Hernán and Robins, 2020).

To control for the confounding factors in an observational study, the propensity score technique for two groups (Rosenbaum and Rubin, 1983) and the generalized propensity score (GPS) for multiple groups (Imbens, 2000) have been widely used. The term GPS refers to the probability of receiving specific treatment assignment conditional on the observed baseline covariates. The alignment of GPS across different treatment groups balances baseline covariates, which approximates the conditional RCT under the exchangeability condition and thus enables us to evaluate ATEs through fairly homogeneous treatment groups. The GPS is often estimated by parametric methods such as multinomial logistic regression, or nonparametric methods such as random forest and generalized boosting methods (McCaffrey et al., 2013). Recently, covariate balancing propensity score method (CBPS) has been proposed to estimate ATE where covariate balancing scores are minimized (Imai and Ratkovic, 2014). Once GPS become available, numerous GPS-based methods (e.g., stratification, matching, inverse probability weighting (IPW), doubly robust method and ensemble doubly robust method) can be used to estimate the ATEs between different treatment groups (Lee et al., 2010; Rosenbaum and Rubin, 1985; Rosenbaum, 1987; Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004; Yan et al., 2019; Hernán and Robins, 2020). However, most of these methods are limited to the pairwise comparisons and may fail to control the family-wise error rate (FWER), leading to a high chance of false discoveries on treatment effects.

To control the FWER, parallel to the well-known Pearson χ^2 test and F test developed in RCT, we develop a weighted χ^2 test for a categorical outcome variable and a weighted F test for a continuous outcome variable to test whether there is

an overall group difference among multiple treatment groups. Only if there is an overall significant group difference, the pairs of interests are further compared. Alternatively, Bonferroni correction is applied to control the FWER for multiple group comparisons. To adjust for the confounding factors, our test procedures first utilize the GPS-based IPW method to create a pseudo population in which the distribution of each confounding factor is similar across different treatment groups (Rosenbaum, 1987; Hernán and Robins, 2020; Robins et al., 2000). We further standardize the weight by a factor so that the total sample size stays the same as the original sample size. Our simulation studies show that the proposed tests can not only control the FWER under the null hypothesis but also achieve great testing power under the alternative hypothesis.

The proposed methods are innovative from three aspects. First, to the best of our knowledge, they are the first effort to test the overall group difference among multiple treatment groups in observational studies. Second, the proposed tests are in alignment with the essence of propensity-score-based method. They can be viewed as the extension of the Pearson χ^2 test for a contingency table and the F test in one-way analysis of variance (ANOVA), respectively. This nature makes it easy to carry out the proposed tests and interpret the results.

The rest of this paper is structured as follows. In section 3.2, we develop a weighted χ^2 test for a categorical outcome variable and a weighted F test for a continuous outcome variable. Section 3.3 presents the simulation studies that are carried out to examine the performance of the proposed tests. Section 3.4 illustrates the usefulness of our proposed tests: we apply the proposed weighted χ^2 test to assess whether fruit/vegetable intakes are associated with heart attack, using the 2015 Kentucky behavioral risk factor surveillance system (BRFSS) dataset, and we apply the weighted F test to examine the effect of physical/recreational exercise on weight gain, using the national health and nutrition examination survey data I

epidemiologic follow-up study (NHEFS) dataset. At last, we conclude the paper with a brief discussion in Section 3.5.

3.2 Weighted χ^2 test and F test

Let X, A, and Y denote, respectively, the vector of p covariates, the treatment received, and the outcome variable in an observational study, where $A \in \{1, \cdots, M\}$ with M being the number of treatment groups ($M > 2$). Moreover, let $Y^{(a)}$ be the potential outcome that would have been observed from a subject under treatment a ($a = 1, \cdots, M$), i.e., there are M potential outcomes for each subject, say $(Y^{(1)}, Y^{(2)}, \cdots, Y^{(M)})$. However, only one potential outcome is observed, which is the outcome corresponding to the treatment actually received, that is, $Y = Y^{(A)}$ (i.e., assuming consistency condition holds). In addition, we assume that (i) there is no unmeasured confounding (i.e., all the confounding variables are measured and included in X, which is also referred as the exchangeability condition), and (ii) the conditional probability for a subject with confounding variable X to be assigned to each treatment group is positive (i.e., the assumption of positivity) (Hernán and Robins, 2020).

Without loss of generality, let's assume that there are N observed independent replicates of (X, A, Y), denoted by $\{(X_i, A_i, Y_i), i = 1, ..., N\}$. We intend to investigate whether there is an overall significant difference in the potential outcome among the M treatment groups. Note that not having an treatment effect is equivalent to that the distribution of $Y^{(a)}$ is the same across the M treatment groups. We can formally write the null hypothesis H_0 and the alternative hypothesis H_1 as follows:

$$\begin{aligned} &H_0 \text{ :the distribution of } Y^{(a)} \text{ does not depend on } a; \\ &H_1 : \text{the distribution of } Y^{(a)} \text{ differs from the distribution of } Y^{(a')}. \end{aligned} \tag{3.1}$$

Since $Y^{(a)}$ is the potential outcome under the treatment a, and not all subjects are assigned to this treatment, the distribution of $Y^{(a)}$ is often unknown. The distribution of observed Y under treatment $A = a$ is often impacted by both the treatment a and the confounding variables X. To carry out the hypothesis test appropriately, the confounding variables must be considered. The IPW method has been a popular and powerful method to estimate ATE and can be used to construct appropriate test statistics while controlling for confounding variables. Under the assumptions of the exchangeability (i.e., there is no unmeasured confounding), positivity, and consistency (Hernán and Robins, 2020), we develop a weighted χ^2 test for categorical outcome variables and a weighted F test for continuous outcome variables to carry out valid hypothesis tests.

To develop the weighted test statistics, we must construct a proper weight for each observation. To this end, we form a pseudo population, where each confounding variable has similar distribution across the M different treatment groups. Given a subject with observed values (x, a, y), the IPW method produces $1/p(a|x)$ many pseudo observations with values (x, a, y), where $p(a|x)$ is the probability of receiving treatment a given the covariates x. The probability vector $\{p(a|x), a = 1, \cdots, M\}$ are often referred to as the GPS (Imbens, 2000; Lunceford and Davidian, 2004). By the law of large number, the number of observations receiving treatment a in the pseudo population is $\sum_{i=1}^{N} I_{\{A_i=a\}}/p(a|X_i) \approx N$, suggesting that each treatment group in the pseudo population simulates the situation that all the subjects in the entire original sample received the intervention. Therefore, the confounding factors do not have any impact on the treatment selection A in the pseudo population, and the pseudo population approximates a RCT (Hernán and Robins, 2020). Note that using the weight $1/p(A_i|X_i)$ $(i = 1, \cdots, N)$ results in a total sample size of the pseudo population as $N_{ipw} := \sum_{a=1}^{M} \sum_{i=1}^{N} I_{\{A_i=a\}}/p(a|X_i)$, which is roughly MN with each treatment group having a sample size N. This motivates us to standardize the

sample size of the pseudo population as follows: for the i^{th} observation (X_i, A_i, Y_i), we generate $NN_{ipw}^{-1}/p(A_i|X_i)$ instead of $1/p(A_i|X_i)$ many pseudo observations, resulting in a total sample size of N as observed but with approximately equal sample size N/M per group and without confounding if the GPS model is correctly specified. Moreover, the observations in the pseudo population can be viewed as independent, and the outcomes of the pseudo observations receiving treatment a are replicates of $Y^{(a)}$ (Lunceford and Davidian, 2004; Hernán and Robins, 2020).

Throughout the paper, we use $*$ to denote the quantities obtained from the pseudo population. For example, $n_a^* = NN_{ipw}^{-1} \sum_{i=1}^{N} I_{\{A_i=a\}}/p(a|X_i)$ denotes the number of subjects receiving treatment a in the pseudo population.

3.2.1 A weighted χ^2 test for categorical outcomes

We first consider the cases with categorical outcomes. Without loss of generality, we assume that Y is a categorical variable with K levels. Let $N_{ak} = \sum_{i=1}^{N} I_{\{A_i=a\}} I_{\{Y_i=k\}}$ denote the number of observations in the treatment group a with outcome variable at k^{th} level, where $a = 1, \cdots, M$ and $k = 1, \cdots, K$. Likewise, let $N_{ak}^* = NN_{ipw}^{-1} \sum_{i=1}^{N} I_{\{A_i=a\}} I_{\{Y_i=k\}}/p(a|X_i)$ denote the counterpart of N_{ak} in the pseudo population in which there is no confounding between treatment assignment and outcome. Then the original observed sample and the pseudo population can be organized into two $M \times K$ contingency tables, as shown in Table 3.1. We use N_{a+}, N_{a+}^* and N_{+k}, N_{+k}^* to denote the row sums and column sums in the two contingency tables, where the subscript "+" denotes the sum over that index.

Table 3.1: The contingency tables based on the observed sample (a) and the pseudo population (b)

(a) The observed sample

	$Y=1$	$Y=2$	$\cdots$	$Y=K$	Total
$A=1$	N_{11}	N_{12}	$\cdots$	N_{1K}	N_{1+}
$A=2$	N_{21}	N_{22}	$\cdots$	N_{2K}	N_{2+}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$A=M$	N_{M1}	N_{M2}	$\cdots$	N_{MK}	N_{M+}
Total	N_{+1}	N_{+2}	$\cdots$	N_{+K}	N

(b) The pseudo population

	$Y=1$	$Y=2$	$\cdots$	$Y=K$	Total
$A=1$	N_{11}^*	N_{12}^*	$\cdots$	N_{1K}^*	N_{1+}^*
$A=2$	N_{21}^*	N_{22}^*	$\cdots$	N_{2K}^*	N_{2+}^*
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$A=M$	N_{M1}^*	N_{M2}^*	$\cdots$	N_{MK}^*	N_{M+}^*
Total	N_{+1}^*	N_{+2}^*	$\cdots$	N_{+K}^*	N

By the new representation of the data from the pseudo population as shown in Table 3.1(b), testing the hypotheses in (3.1) is equivalent to testing whether the column variable (i.e., outcome variable) is independent of the row variable (i.e., treatment variable). Since there are no confounding factors between treatment and outcome variable in Table 3.1(b), we can apply the Pearson χ^2-test. Let $E^*_{ak} = N^*_{a+}N^*_{+k}/N$ be the expected count in the cell $(A = a, Y = k)$. Then the weighted test statistic, $w\chi^2$, can be computed as

$$w\chi^2 = \sum_{k=1}^{K}\sum_{a=1}^{M}\frac{(N^*_{ak} - E^*_{ak})^2}{E^*_{ak}}. \tag{3.2}$$

Under the null hypothesis that there is no treatment effect (i.e., the distribution of outcome is same across different treatment groups), the test statistic $w\chi^2$ follows a χ^2 distribution with $(M-1)(K-1)$ degrees of freedom. Larger values of $w\chi^2$ lead to more evidence to reject H_0.

If H_0 is rejected, we can follow up to calculate the standardized Pearson residuals e_{ak}'s (Haberman, 1973; Agresti, 2012):

$$e_{ak} = \frac{N^*_{ak} - E^*_{ak}}{\left[E^*_{ak}\left(1 - N^*_{+k}/N\right)\left(1 - N^*_{a+}/N\right)\right]^{1/2}},$$

which may provide additional information about the causal effect of the treatment a on the potential outcome $Y^{(a)}$. A larger e_{ak} indicates that treatment a results in a larger proportion of response at k^{th} level, that is, treatment a favors response at k^{th} level. More rigorous inference for two group comparison (say group a_1 versus a_2) can

be made using the χ^2 test statistic:

$$w\chi^2(a_1, a_2) = \sum_{k=1}^{K} \sum_{a \in \{a_1, a_2\}} \frac{(N^*_{ak} - E^*_{ak})^2}{E^*_{ak}}, \tag{3.3}$$

which follows a χ^2 distribution with $K - 1$ degrees of freedom. The comparisons between any two treatments are carried out only if the hypothesis test for H_0 against H_1 in (3.1) is rejected. Thus, the FWER is controlled. Alternatively, we can use Bonferroni correction method to control FWER without carrying out the overall weighted χ^2 test in equation (3.2). That is, we carry out the comparison between any two treatments a_1 versus a_2 as in equation (3.3) using Bonferroni adjusted p-values.

3.2.2 A weighted F test for continuous outcomes

We next consider the cases with continuous outcomes. One plausible way to test the hypotheses in (3.1) is to group the outcomes into a small number of categories and then apply the proposed χ^2 test in Section 3.2.1. However, there is usually no clear scientific guideline for how to group the data, and different ways of grouping may lead to different conclusions. Thus, we instead construct weighted F statistics to test whether there is an overall treatment difference among multiple treatment groups for continuous outcomes.

Let $\mu_a = E[Y^{(a)}]$ denote the population mean given that the entire population has been assigned to treatment a. We test

$$H'_0 : \mu_1 = \cdots = \mu_M \text{ against } H'_1 : \mu_a \neq \mu_{a'} \text{ for some } a, a' \in \{1, \cdots, M\}. \tag{3.4}$$

It is easy to see that the hypothesis H'_0 in (3.4) is weaker than H_0 in (3.1) for continuous outcomes. The null hypothesis H'_0 intuitively indicates that the mean outcome in the entire population receiving each treatment remains the same, while the alternative hypothesis H'_1 implies that the mean outcomes are different at least under two

different treatment groups.

Assume that $Y^{(a)}$ follows a normal distribution and the variances of $Y^{(a)}$ are the same for all a's. Since the confounding factors do not impact the selection of treatment in the pseudo population created by the IPW, we can apply the F test for one-way ANOVA to test H_0' against H_1' using the pseudo population. The population mean μ_a under treatment a can be simply estimated by the sample average of the outcomes of subjects receiving treatment a in the pseudo population:

$$\hat{\mu}_a^* = \frac{NN_{ipw}^{-1}\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}Y_i}{n_a^*} = \frac{\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}Y_i}{\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}}.$$

Likewise, taking the average of all outcomes in the pseudo population yields an estimate of the grand mean:

$$\hat{\mu}^* = \frac{\sum_{a=1}^{M}\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}Y_i}{\sum_{a=1}^{M}\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}}.$$

Consequently, the sum of squares for treatments (SST^*) and the sum of squares for errors (SSE^*) for the pseudo population can be obtained as follows:

$$SST^* = \sum_{a=1}^{M} n_a^*(\hat{\mu}_a^* - \hat{\mu}^*)^2, \quad SSE^* = NN_{ipw}^{-1}\sum_{a=1}^{M}\sum_{i=1}^{N} 1/p(a|X_i)I_{\{A_i=a\}}(Y_i - \hat{\mu}_a^*)^2.$$

The weighted F statistic testing H_0' against H_1' based on the pseudo population can be obtained as follows:

$$wF = \frac{\text{variance between treatments}}{\text{variance within treatments}} = \frac{SST^*/(M-1)}{SSE^*/(N-M)}. \tag{3.5}$$

Assume that the outcomes in the pseudo population are independent and normally

distributed with the same variance, the test statistic wF follows a F distribution with degrees of freedom $M-1$ and $N-M$ under H_0'. Table 3.2 summarizes the variation sources in the pseudo population.

Table 3.2: Source of variation for the pseudo population in the proposed weighted F test for continuous outcomes.

Source of variation	DF	Sum of squares	Mean squares	wF test statistic
Between groups	$M-1$	SST^*	$MST^* = SST^*/(M-1)$	$wF = \frac{MST^*}{MSE^*}$
Within groups	$N-M$	SSE^*	$MSE^* = SSE^*/(N-M)$	
Total	$N-1$	SS^*		

A larger value of wF leads to more evidence to reject H_0'. If H_0' is rejected, the pairwise comparison between treatments a and a' can be further conducted using a weighted student t test:

$$wt(a,a') = \frac{\hat{\mu}_a^* - \hat{\mu}_{a'}^*}{\sqrt{MSE^*(1/n_a^* + 1/n_{a'}^*)}}, \tag{3.6}$$

which follows a central t distribution with $n_a^* + n_{a'}^* - 2$ degrees of freedom under the null hypothesis that $\mu_a = \mu_{a'}$. The MSE^* in (3.6) and Table 3.2 is an estimate of the within-group variance. Since we perform the weighted t test only if we reject H_0' using the weighted F test, the FWER is controlled. Alternatively, we can use Bonferroni correction to control FWER without performing the overall weighted F test. That is, we conduct the group comparison between treatment a versus a' using equation (3.6) with Bonferroni corrections.

As the GPS $\{p(a|X_i), a = 1, \cdots, M, i = 1, \cdots, N\}$ are unobserved in practice, our proposed test statistics $w\chi^2$ in (3.2) and wF in (3.5) are computed with replacing the GPS by their estimates of $\{\hat{p}(a|X_i), a = 1, \cdots, M, i = 1, \cdots, N\}$. If the estimates of the GPS are consistent, the proposed test statistics are still valid with the same distributions, by the Slutcky's theorem (Casella and Berger, 2002).

Traditionally, GPS is estimated using a multinomial regression model (Imbens, 2000). Recently, Imai and Ratkovic (2014) developed a CBPS method to estimate

GPS and balance covariates among different groups simultaneously. In our simulation study, both traditional GPS estimation and CBPS estimation are investigated to examine their performance in estimating ATEs and test the treatment effects.

3.3 Simulation study

We conduct simulation studies to examine the performance of the proposed tests. In order to carry out the proposed tests, we obtain the estimates of GPS using the multinomial logistic regression model and the CBPS method (Imai and Ratkovic, 2014). In this simulation study, we also construct the weighted χ^2 test and weighted F test using the true GPS. We compare the proposed tests to their counterparts that do not adjust for the confounding factors: the Pearson χ^2 test for categorical outcomes and F test for continuous outcomes.

3.3.1 Simulation settings

In the simulation study, we set up four confounding variables, say $X = (X_1, X_2, X_3, X_4)'$. We set up three treatment groups, that is, $M = 3$. The treatment assignments A are generated from a multinomial distribution that depends on the confounding variable X. We consider two types of outcome variables: a categorical outcome in Scenario I and a continuous outcome in Scenario II.

- **Scenario I:** The outcome Y is generated from the logistic regression model:

$$\ln\left(\frac{\Pr(Y=1|X,A)}{1-\Pr(Y=1|X,A)}\right) = X'\alpha + 0.5\tau I_{\{A=2\}} + \tau I_{\{A=3\}}.$$

  where $\alpha = (0.125, 2.10, 1.25, 1.50)'$, and τ captures the treatment effect: the odds of the outcome Y being 1 from a subject receiving treatment 2 is $\exp(0.5\tau)$ times of that from a subject receiving treatment 1, and the odds of the outcome Y being 1 from a subject receiving treatment 3 is $\exp(\tau)$ times of that from a

subject receiving treatment 1. $\tau = 0$ represents the situation that there is no group difference among the three treatment groups.

- **Scenario II:** The outcome Y is generated from the multiple linear regression model:

$$Y = -0.8X_1 + 2X_2 + 2X_3 + 2X_4 + 0.5\tau I_{\{A=2\}} + \tau I_{\{A=3\}} + \varepsilon.$$

where ε follows a standard normal distribution. The true ATE is 0.5τ between treatments 1 and 2, and the true ATE is τ between treatments 1 and 3. Once again, $\tau = 0$ indicates that there is no group difference among the three groups.

We choose τ as an equally-spaced sequence from 0 to 5 by step 0.2 to examine both size and power of different tests. We set the sample size N to 100, 500 and 1000, respectively. For each τ and each sample size N, we generate 1000 samples. For each sample, we carry out the hypothesis test using different methods. The simulation procedures are described as follows:

Step 1 Generate the vector of covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ for $i = 1, \cdots, N$, where $X_{i1} = 1$ as a constant covariate, X_{i2} follows a standard normal distribution, X_{i3} follows a uniform distribution on the interval $(-0.5, 0.5)$, and X_{i4} is a random variable taking values ± 0.5 with probability 0.5 for each value.

Step 2 Generate treatment assignment variable $A_i \in (1, 2, 3)$ $(i = 1, \cdots, N)$. Given X_i, the treatment A_i follows a multinomial distribution with the following parameter:

$$\Pr(A_i = a|X_i) = \frac{\exp(X_i'\beta^{(a)})}{\sum_{k=1}^{3}\exp(X_i'\beta^{(k)})}, \quad a = 1, 2, 3. \tag{3.7}$$

where $\beta^{(1)} = (0, 0, 0, 0)'$, $\beta^{(2)} = (-0.2, 0.15, 0.2, 0.1)'$, and $\beta^{(3)} = (-0.25, 0.3, 0.4, 0.2)'$.

Step 3 Generate outcome variables based on the model in Scenario I for the categorical outcome and the model in Scenario II for the continuous outcome.

Step 4 Carry out the hypothesis tests using χ^2 (or F) test, the proposed weighted $w\chi^2$ (or wF) test with true GPS, estimated GPS using multinomial logistic regression, and estimated GPS using CBPS method, respectively. A test is significant if the p-value is less than 0.05.

Step 5 Carry out the hypothesis tests to compare all pairs using χ^2 (or t) test, $w\chi^2$ (or wt) test with true GPS, estimated GPS using multinomial logistic regression, and estimated GPS using CBPS method, respectively. The p-values are adjusted using Bonferroni correction, and the comparisons between any two treatment groups are carried out for each method. We make a decision on whether we reject any one of the comparisons for each method, thus we enable to evaluate the FWER. Here the significance level is set as 0.05.

Step 6 Repeat Steps 1-5 1000 times, and summarize the rejection rate among the 1000 simulated datasets for each method.

Step 7 Repeat Steps 1-6 for each fixed τ, where τ is a sequence from 0 to 5 by step 0.2.

The simulation results are reported in Figure 3.1 for $N = 100$, Figure 3.2 for $N = 500$, and Figure A1.7 for $N = 1000$. In each figure, we report the family-wise rejection rates versus different τ for each statistical method mentioned in Steps 4 and 5. When the outcome variable is categorical (Scenario I), for a fixed τ and sample size N, we calculate the rejection rates among the 1000 generated data for each of the methods, which include the traditional χ^2 test and weighted $w\chi^2$ for testing whether the treatment assignment is independent of the outcome (see methods in Step 4 and results in Panel A in each figure), and the traditional χ^2 test and weighted $w\chi^2$ for testing whether there is a group difference between two treatment groups adjusted by

Bonferroni correction (see methods in Step 5 and results in Panel B in each figure). When the outcome variable is continuous (Scenario II), for a fixed τ and sample size N, we calculate the rejection rates among the 1000 generated data for each of the methods, which include the traditional F test and weighted wF for testing whether there is a global treatment effect among the three treatment groups (see methods in Step 4 and results in Panel C in each figure), and the student t test and weighted wt test for testing whether there is a pair of treatments whose outcomes are significantly different with Bonferroni correction (see methods in Step 5 and results in Panel D in each figure).

Because the power of a test is defined as the probability of rejecting the null hypothesis under the alternative hypothesis, the curves for the rejection rates versus τ are referred as power curves in Figures 3.1, 3.2, and A1.7. The rejection rates at $\tau = 0$ are referred as type I error rates. We run the simulation study for τ from 0 to 5 for all different sample sizes, the power of all the tests are already close to 1 at τ=3 when sample sizes are 500 and 1000. Thus, we only present the results for τ from 0 to 3 in Figures 3.2 and A1.7.

3.3.2 Simulation results

In the simulation study, the weights used in the weighted tests are obtained from the true GPS (see the dashed lines in Figures 3.1, 3.2, and A1.7), estimated GPS using a multinomial logistic regression (see the dotted lines), and estimated GPS using the CBPS method (see dash-dotted lines). Based on the simulation results in Figures 3.1, 3.2, and A1.7, we draw the following conclusions:

1. The proposed weighted tests can successfully control the overall FWER, that is, the rejection rate when $\tau = 0$. For categorical outcome variables, it can be seen that the weighted χ^2 test with the estimated GPS (wχ^2.MLR.PS or wχ^2.CBPS.PS) could successfully control the rate of the type I error below

5% in Figures 3.1A, 3.2A and A1.7A. For continuous outcome variables, it is clear that the weighted F test with the estimated GPS (wF.MLR.PS and wF.CBPS.PS) have a satisfactory FWER based on the panel C in Figures 3.1, 3.2, and A1.7.

2. By adjusting the p values with Bonferroni correction, the pairwise group comparisons with estimated GPS could control FWER as well (see wχ^2.MLR.PS and wχ^2.CBPS.PS for categorical outcome variables in Figures 3.1B, 3.2B and A1.7B and wt.MLR.PS and wt.CBPS.PS for continuous outcome variables in Figures 3.1D, 3.2D and A1.7D).

3. The traditional tests including the χ^2 test, F test, the pairwise χ^2 test with Bonferroni correction, and the pairwise t test with Bonferroni correction (see the solid line in each panel), obviously fail to control the FWER for either type of outcome variables. Therefore, the traditional tests are not appropriate for testing treatment effect when there is confounding.

4. From Figures 3.1, 3.2 and A1.7, we also observe that the power of each weighted test increases as τ increases. Among all the weighted tests (i.e., true GPS, estimated GPS using a multinomial logistic regression model, and estimated GPS using CBPS), the weighted tests with the consistently estimated GPS are comparable or better than the tests with the true GPS. Moreover, we note that the weighted tests with GPS estimated by CBPS are superior to the tests with GPS from a multinomial logistic regression model. This finding is in alignment with the merit of the CBPS method, which estimates GPS with consideration of balancing the covariates (Imai and Ratkovic, 2014).

5. By comparing Figures 3.1, 3.2 and A1.7, it is clear that as the sample size increases, teach weighted test increases, while the rate of type I error is still well controlled about 0.05. For example, the power of each weighted test with

sample size 100 is about 20% at τ=1 (Figure 3.1), while the power of each weighted test is greater than 60% for sample size 500 at τ=1 (Figure 3.2), and reaches almost 1 for sample size 1000 at τ=1 (Figure A1.7). However, when the sample size increases, the inflation of the type I error of the traditional tests becomes even larger.

In the following case studies, we use the weighted tests with GPS estimated using the CBPS method.

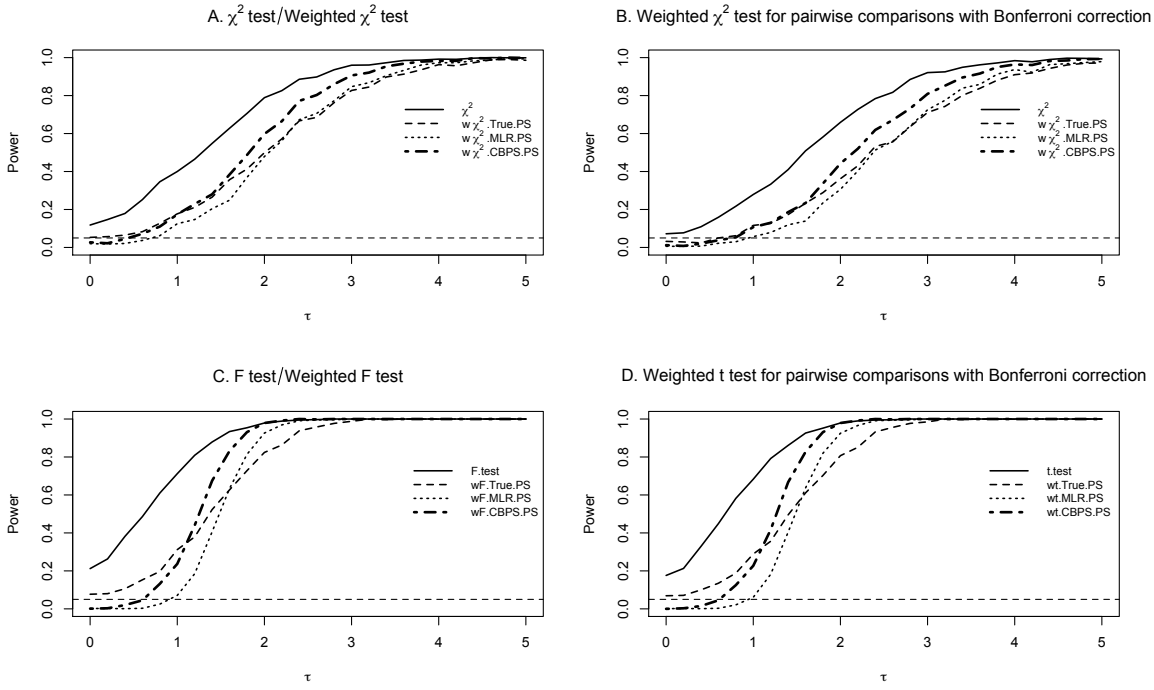

Figure 3.1: Power curves of different tests with sample size 100. In each panel, the solid line represents the traditional test, the dashed line represents the weighted test using the true GPS, the dotted line represents the weighted test using GPS estimated by multinomial logistic regression (MLR) model, and the dash-dotted line represents the weighted test with GPS estimated using CBPS method. The horizontal line is at a height 0.05, the nominal size of the test.

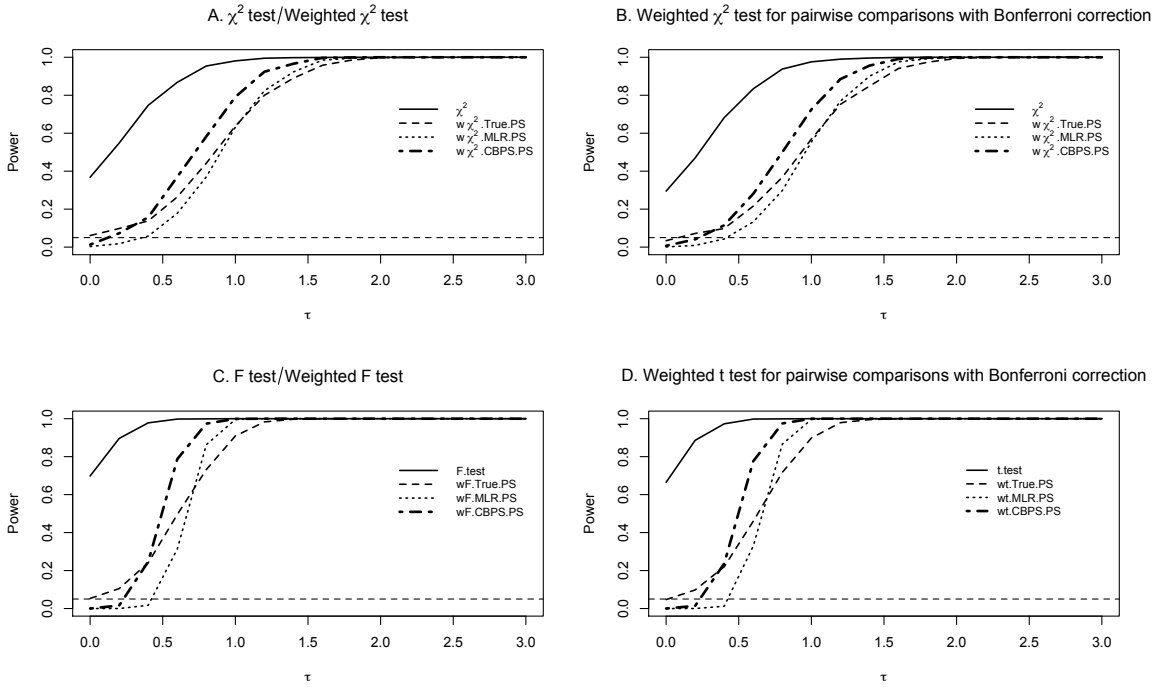

Figure 3.2: Power curves of different tests with sample size 500. In each panel, the solid line represents the traditional test, the dashed line represents the weighted test using the true GPS, the dotted line represents the weighted test using GPS estimated by multinomial logistic regression (MLR) model, and the dash-dotted line represents the weighted test with GPS estimated using CBPS method. The horizontal line is at a height 0.05, the size of the test.

3.4 Case studies

In this section, we use two data sets to illustrate the practical usage of the proposed tests. We examine the effect of healthy diet on heart attack using the 2015 Kentucky behavioral risk factor surveillance system (BRFSS) dataset, and we also examine the impact of physical exercise on weight gain using the first national health and nutrition examination survey (NHANES I) epidemiologic follow-up study (NHEFS) dataset.

3.4.1 Study healthy diet on heart attack using 2015 Kentucky BRFSS dataset

The BRFSS is the national premier system of health-related telephone surveys that collect data about U.S. residents at each state regarding their health-related risk

behaviors, chronic health conditions, and use of preventive services. In this case study, we apply the proposed weighted χ^2 test to the 2015 Kentucky BRFSS data to investigate whether the healthy diet impacts cardiovascular diseases, in particular, heart attack. There are 111,379 subjects included in this dataset.

Heart attack is the major cardiovascular disease (CVD). Several studies showed possible protective effects of healthy diet and physical activities on the CVD (Moore et al., 2015; Dietary Guidelines Advisory Committee, 2015; CDC, 2013). We would like to examine the effect of healthy diet on heart attack using the proposed test. We consider the following confounding variables: education, age, gender, race, median income, and percentage of below poverty determined by the zip code level where the subject lives. According to the national guidelines on fruit/vegetable consumption provided by the American College of Sports Medicine and the Centers of Disease Control and Prevention (Dietary Guidelines Advisory Committee, 2015; Dauchet et al., 2009; WHO, 2003): "adults should consume fruits and vegetables 5 cups or more times daily or consume fruit 2 or more cups and vegetables 3 or more cups daily", we classify all the subjects into three groups: (1) G_1-Neither: neither vegetable nor fruit consumption meets guide lines; (2) G_2-Either: either vegetable or fruit consumption meets guide lines; (3) G_3-Both: both vegetable and fruit consumption meet guide lines. We consider the outcome variable heart attach (Yes/No). For the continuous variables in the data, we summarize their information by calculating mean and standard error, while for the categorical variables, the summary information is provided in terms of counts and percentages, stratified by the three treatment groups. The summary information is shown under "Observed sample" in Table 3.3.

Before testing the treatment effect, we need to estimate the propensity score and the weight for each observation to form the pseudo population, and then construct the weighted χ^2 test statistics to test the causal effect of fruit/vegetable consumption on heart attack. We also summarize the information of the related variables in the

pseudo population (see "Pseudo population" in Table 3.3). It is clear that the distribution of the covariates under dfferent groups in the pseudo population are similar. We use the weighted χ^2 test to examine whether there is an overall significant difference among the three treatment groups. Based on the test result, we conclude that there is overall significant group difference. The follow-up weighted pairwise χ^2 test indicates that the difference is significant for any two pairs among the three groups. The test results together suggest that the healthy diet has a significant protective effect on heart attack.

Table 3.3: The summary of the variables under the three diet groups in the observed sample and pseudo population

		Observed sample			Pseudo population		
		G_1-Neither	G_2-Either	G_3-Both	G_1-Neither	G_2-Either	G_3-Both
	Sample size	80,999	24,084	6,296	37,174	37,035	37,170
		(72.7%)	(21.6%)	(5.7%)	(33.4%)	(33.2%)	(33.4%)
Covariates	**Median Income**	44,543	45,374	46,388	44,824	44,797	44,633
	% of below poverty	20.2%	20.1%	19.7%	20.2%	20.2%	20.2%
	Education						
	< HS	7534	1712	216	3145	3189	4447
		(9.3%)	(7.1%)	(3.4 %)	(8.5%)	(8.6%)	(12.0%)
	HS	26276	7080	989	11436	11468	10987
		(32.4%)	(29.4%)	(15.7%)	(30.8%)	(31.0%)	(29.6%)
	College	23393	7028	1582	10782.0	10593.8	9376.7
		(30.0%)	(29.2%)	(25.1%)	(29.0%)	(28.6%)	(25.2%)
	Graduate	23796	8264	3509	11810.1	11783.7	12359.3
		(29.4%)	(34.3%)	(55.7%)	(31.8%)	(31.8%)	(33.3%)
	Age	56	57	55	56.3	56.1	57.1
	Gender						
	Male	30471	7463	1215	13064.8	13058.3	12581.0
		(37.6%)	(31.0%)	(19.3%)	(35.1%)	(35.3%)	(33.8%)
	Female	50528	16621	5081	24109.1	23976.7	24589.0
		(62.4%)	(69.0%)	(80.7%)	(64.9%)	(64.7%)	(66.2%)
	Race						
	White	60871	17238	4350	27837.0	26761.7	26055.4
		(75.2%)	(71.6%)	(69.1%)	(74.9%)	(72.3%)	(70.1%)
	Non-white	20128	6846	1946	9337.0	10273.3	11114.6
		(24.9%)	(28.4%)	(30.9%)	(25.1%)	(27.7%)	(29.9%)
Outcome	**Heart attack**$\star$					ⓐ	ⓐⓡ
	Yes	6592	1716	255	2955.5	2642.4	2247.5
		(8.1%)	(7.1%)	(4.0%)	(8.0%)	(7.1%)	(6.0%)
	No	74407	22368	6041	34218.5	34392.7	34922.5
		(91.9%)	(92.9%)	(96.0%)	(92.0%)	(92.9%)	(94.0%)

Note: $\star$ indicates that there is significant difference among the three treatment groups; ⓐ indicates a significant difference from group G_1-Neither; ⓡ indicates a significant difference from group G_2-Either.

3.4.2 Study physical exercise on weight gain using the NHEFS dataset

The NHANES I epidemiology follow-up study (NHEFS) is a national longitudinal study that was jointly initiated by the national center for health statistics and the national institute on aging in collaboration with other agencies of the public health service. The NHEFS was designed to investigate the relationships between clinical,

nutritional, and behavioral factors assessed in the NHANES I and subsequent morbidity, mortality, and hospital utilization, as well as changes in risk factors, functional limitation, and institutionalization. We merged the NHEFS 1982 data with NHANES I to investigate the effect of the physical/recreation exercise on the weight gain from 1971 to 1982. The combined data consists of 2842 subjects, who are divided into three groups based on their physical/recreation exercises: (1) G_1-Inactive; (2) G_2-Moderate; (3) G_3-Intensive. We assume that the following eight baseline variables are sufficient to adjust for confounding: gender, age, race, education, diet, smoking, income and weight in 1971 in pound. The outcome is the weight gain from 1971 to 1982. Table 3.4 presents the summary statistics of the eight baseline covariates and the outcome (i.e., weight gain) among three treatment groups before and after adjusting for the confounding factors. The continuous variables are summarized by mean and standard error, and the categorical variables are summarized by counts and percentages, stratified by three groups, which are shown under "Observed sample". The column under the "Outcome" in Table 3.4 is the summarized mean and standard error of the weight gain for each level of a categorical variable, or the regression slope of weight gain on the continuous covariate, where $\diamondsuit$ indicates that the slope is significantly different from zero. For example, the percentage of female in G_3-Intensive group is lower than that in the other two groups, and the weight gain in female is smaller than that in male. The subjects in G_3-Intensive group are younger than the subjects in the other two groups, and the weight gain is negatively associated with age (slope -0.314 with p-value¡0.05). Thus age and gender could be confounding factors for physical/recreation exercises and weight gain. We use the IPW method to create a pseudo population to remove the impact of these confounding factors. The summary statistics of the pseudo population are reported under "Pseudo population" in Table 3.4. It is clear that the covariates are similar among the three groups in the pseudo population. We apply the weighted F test to test the overall treatment effect, and p-value 0.015 suggests that

the weight gain is significantly different among these three groups. Subsequently, we conduct pairwise comparisons by using the weighted t test. The test results indicate that there is no significant difference on weight gain between groups G_1-Inactive and G_2-Moderate or between G_2-Moderate and G_3-Intensive. However, the weight gain is significantly different between G_1-Inactive and G_3-Intensive (p-value 0.015). Thus, we conclude that the physical/recreation exercises impact the weight gain, and the intensive exercise tends to gain more weight than inactive group.

Table 3.4: The summary of the variables stratified by groups in the observed sample as well as in the pseudo population

	Observed sample				Pseudo population		
Covariates	G_1-Inactive	G_2-Moderate	G_3-Intensive	Outcome	G_1-Inactive	G_2-Moderate	G_3-Intensive
	189	1274	1379		942	951	949
	(6.7%)	(44.8%)	(48.5%)		(33.1%)	(33.5%)	(33.4%)
Gender							
Male	84	660	839	5.1	521.7	529.5	528.8
	(44.4%)	(51.8%)	(60.8%)	(0.4)	(55.4%)	(55.7%)	(55.7%)
Female	105	614	540	3.7	420.3	421.5	420.2
	(55.6%)	(48.2%)	(39.2%)	(0.5)	(44.6%)	(44.3%)	(44.3%)
Age	48.2	47.2	44.8	-0.314◇	46.4	46.1	46.1
	(0.9)	(0.4)	(0.3)		(0.4)	(0.4)	(0.4)
Race							
White	163	1127	1241	4.4	835.4	847.9	845.5
	(86.2%)	(88.5%)	(90.0%)	(0.3)	(88.7%)	(89.2%)	(89.1%)
Non_white	26	147	138	4.6	106.6	103.1	103.5
	(13.6%)	(11.5%)	(10.0%)	(1.3)	(11.3%)	(10.8%)	(10.9%)
Education							
< HS	43	244	261	1.3	172.5	182.4	181.1
	(22.8%)	(19.2%)	(18.9%)	(0.8)	(18.3%)	(19.2%)	(19.1%)
HS drop	34	210	260	5.4	169.7	166.5	172.3
	(18%)	(16.5%)	(18.9%)	(0.8)	(18%)	(17.5%)	(18.2%)
HS	63	465	523	5.5	356.1	350.8	350.6
	(33.3%)	(36.5%)	(37.9%)	(0.5)	(37.8%)	(36.9%)	(36.9%)
College	31	279	265	4.5	186.4	194.3	189.7
	(16.4%)	(21.9%)	(19.2%)	(0.7)	(19.8%)	(20.4%)	(20%)
Graduate	18	76	70	4.9	57.3	56.9	55.3
	(9.5%)	(6%)	(5.1%)	(1.3)	(6.1%)	(6.0%)	(5.8%)
Diet							
Yes	35	170	128	3.3	108.5	110.7	109.1
	(18.5%)	(13.3%)	(9.3%)	(1.1)	(11.5%)	(11.6%)	(11.5%)
No	154	1104	1251	4.6	833.5	840.3	839.8
	(81.5%)	(86.7%)	(90.7%)	(0.3)	(88.5%)	(88.4%)	(88.5%)
Smoking							
Yes	116	755	867	5.6	585.5	572.4	583.5
	(61.4%)	(59.3%)	(62.9%)	(0.4)	(62.2%)	(60.2%)	(61.5%)
No	73	519	512	2.6	356.5	378.7	365.5
	(38.6%)	(40.7%)	(37.1%)	(0.5)	(37.8%)	(39.8%)	(38.5%)
Income							
< $6,000	60	294	262	2.1	206.1	205.2	204.2
	(31.7%)	(23.1%)	(19%)	(0.8)	(21.9%)	(21.6%)	(21.5%)
($6,000, $20,000)	104	774	922	5.2	597.2	601.8	602.6
	(55%)	(60.8%)	(66.9%)	(0.4)	(63.4%)	(63.3%)	(63.5%)
>$20,000	25	206	195	4.5	138.7	144	142.2
	(13.2%)	(16.2%)	(14.1%)	(0.7)	(14.7%)	(15.1%)	(15.0%)
Weight_1971	169.9	162.1	158.5	-0.084 ◇	159.6	160.8	160.5
	(3.5)	(1.0)	(0.9)		(1.2)	(1.1)	(1.1)
Weight gain ⋆	-1.1	3.7	5.9		2.6	4.3	5.0@
	(1.7)	(0.5)	(0.4)		(0.6)	(0.5)	(0.6)

Note: ⋆ indicates that there is significant difference among the three treatment groups; @ indicates a significant difference from group G_1-Inactive; ◇ indicates a significant regression coefficient by regressing weight gain on a continuous variable.

3.5 Discussion

In this study, to test whether there is an overall difference among multiple treatment groups in observational study, we propose a weighted χ^2 test for categorical outcomes and a weighted F test for continuous outcomes. The proposed tests are able to make valid inference for group differences by removing the confounding factors between outcome and treatment groups. The simulation results showed that the weighted tests could successfully control the FWER, while the traditional tests without adjusting for the confounding factors had an inflated type I error rate, which means that the traditional tests are not appropriate for testing treatment effect in the observational study. In addition, we used the multinomial logistic model and CBPS methods to estimate the GPS, the test with GSP estimated by CBPS method generally performs better in terms of the power of the test.

Our proposed methods are very intuitive and easy to implement in observational studies. First, IPW method is used to create a pseudo population in which the confounding factors are balanced, and the sample size of the pseudo population is standardized to the same as the original sample size. Subsequently, under the assumptions that the subjects in the pseudo population are independent, the weighted tests are proposed to conduct the global hypothesis test and then conduct the pairwise comparison if there is an overall significant group difference. The FWER is controlled at the specified significance level, say 0.05. If the global hypothesis test is not performed, Bonferroni corrections for group comparisons are also able to control FWER.

Our research work can be expanded from the following three perspectives. First, we could apply the proposed methods to other data type, such as censored survival data and missing data, and examine whether the similar approach can be carried forward. Second, in the current work, we only use the parametric model (i.e.,

multinomial logistic model and CBPS method) to estimate GPS, assuming that the parametric model is correctly specified. However, the parametric model may not be specified correctly. We will investigate non-parametric methods, such as random forests and generalized boosted model (Rubin, 2004; McCaffrey et al., 2013), to obtain more accurate and robust GPS estimates, thus to improve the performances of the weighted test statistics. Last but not least, we will investigate the performance of the stabilized weights, which could be mathematically written as $w(a; X_i) = p(a)/p(a|X_i)$, where $p(a)$ is the marginal probability of receiving treatment a. The stabilized method may narrow the range of the weight, thus to alleviate the effect of the subject with too large weight (Hernán and Robins, 2020).

CHAPTER 4

ESTIMATION OF AVERAGE TREATMENT EFFECT FOR TIME DEPENDENT OUTCOMES

4.1 Introduction

In an observational study, propensity score takes an important role in estimating average treatment effect (ATE) by adjusting for the confounding factors (Rosenbaum and Rubin, 1983). Given the propensity score, a multiple of methods have been proposed to estimate ATE, such as, matching, stratification, inverse probability of treatment weighting (IPTW) and doubly robust methods (Rosenbaum, 1987; Rosenbaum and Rubin, 1984, 1985; Lunceford and Davidian, 2004; Lee et al., 2010; Hernán and Robins, 2020). However, in practice, the true propensity score is unknown, and the performance of these propensity-score-based methods depends on the estimation of the propensity scores. If the propensity score estimation model is correctly specified, these methods could provide consistent estimates of ATE. It is common that researchers include all the covariates into the propensity score estimation model, which is called the "throw in the kitchen sink" approach (Shortreed and Ertefaie, 2017). This method may fail if the number of variables is large (Brookhart et al., 2006). Literature has suggested that including the variables only related to the treatment in the propensity score model may increase the variation of the ATE estimation; however including the variables related to the outcome in the propensity score model may lead to a more accurate ATE estimation (Brookhart et al., 2006; De Luna et al., 2011; Patrick et al., 2011). Thus, variable selection is important in

estimating propensity score, especially when there are a large number of covariates.

As to how to select covariates for the propensity score model, the regularization methods have been applied (Zou, 2006; Shortreed and Ertefaie, 2017). For example, the outcome-adaptive lasso method applies the penalty on the likelihood of the propensity score model, where the tuning parameter is selected to balance the covariates, and the penalty weight for each covariate is the inverse of its absolute regression coefficient in the outcome regression model. Thus, the covariate which is weakly or unrelated to the outcome has a larger penalty, thus forcing the variable out of propensity score estimation model (Zou, 2006; Shortreed and Ertefaie, 2017). Ertefaie et.al. also proposed to select the variables by penalizing simultaneously the outcome model and the treatment assignment model, and showed that the proposed method achieves the oracle properties (Ertefaie et al., 2018). A similar method based on the lasso is used by Franklin et al. (Franklin et al., 2015). Bayesian methods are also used to select confounding variables when the number of covariates is large and the sample size is small (Wang et al., 2015). Zigledepenr and Dominici use Bayesian method to select the variables and obtain the weighted average of the treatment estimates under different propensity score models (Zigler and Dominici, 2014). However, all these variable selection methods are only applied to continuous or categorical outcomes. We extend this technique to time dependent outcomes, such as survival time or the life-time cost since the diagnosis of a disease.

Although a lot of work have been carried out to estimate ATE for survival outcome (Lin and León, 2017; Xu et al., 2012; Xie and Liu, 2005; Austin, 2013, 2014; Austin and Schuster, 2016), the performance of variable selection has not been considered. Three measures are often used to gauge the ATE: (1) the difference between mean or median survival time; (2) absolute reduction of the probability of the occurrence of an event at a certain time point; and (3) hazard ratio (Austin and Schuster, 2016). Many propensity score based methods (e.g., matching, stratifica-

tion, IPTW and doubly robust method) have been extended to time-to-event data. There are two commonly methods to estimate ATE under the framework of IPTW (Austin, 2013, 2014): the weighted Kaplan-Meier estimation and the weighted Cox proportional model, where the weights are the inverse of the probability of treatment received, which is often obtained from the propensity score model. Currently, these two methods are widely applied to survival data, assuming that the survival time and censoring time are independent (Lin and León, 2017; Xu et al., 2012; Xie and Liu, 2005). However, when the censoring time is informative, that is, the censoring may depend on the covariates and treatment, the censoring must be accounted for. For informative censoring, the inverse probability of uncensoring weighting method is applicable (Robins and Rotnitzky, 1992; Robins, 1993), where the weight is the inverse of the probability of uncensoring. The probability of uncensoring can be estimated parametrically or non-parametrically. By weighting each uncensored subject, we create a pseudo population in which all the subjects are completely observed (Schaubel and Wei, 2011), and this pseudo population is similar to the original observed sample in baseline characteristics but without censoring observations. Further, Jiaqi and her college proposed the doubly robust method for comparing medical costs on treatment effect, using the double weights for each uncensored subject (Li et al., 2016).

In this study, we investigate the impact of variable selection on estimating ATE for time dependent outcomes. We first select the covariates which are associated with outcomes, and then use the selected variables to estimate the propensity scores. To remove the impact of the confounding factors and remove the selection bias due to the informative censoring, we propose the doubly weighting method to estimate ATE. One weight is the inverse of the probability of treatment received, and the other one is the inverse of the probability of remaining uncensored. By using the double weights, we create a pseudo population in which all the confounding factors are balanced and all the subjects are uncensored. The innovation is that we use the variable selection

technique to estimate the propensity scores and the probability of uncensoring. We anticipate that the proposed method provides more accurate ATE estimates than its counterpart but wihtout variable selection. We also anticipate that the proposed method is suitable for high dimensional data. In the simulation study, we examine the performance of the variable selection doubly weighting method based on two different censoring type: uniform censoring and informative censoring. Under each censoring type, we compare the performance of our proposed method with the IPTW method, and the no-variable-selection doubly weighting method, varying the number of the covariates. The simulation results show that the doubly weighting method performs better than other methods in terms of unbiasedness and variation, particularly, when the number of covariates is large.

The rest of the paper is structured as followings. In Section 4.2, we develop the variable-selection doubly weighting method to estimate ATE in observational study with informative censoring; In Section 4.3, simulation studies are carried out to examine the performance of the proposed method; In section 4.4, SEER-Medicare data is used to compare the mean survival time of different treatments on pancreas patients. At last, we conclude the paper with a brief discussion in Section 4.5.

4.2 Method

Let assume that we have a quartet of variables (X, A, T_d, Y) for each subject, where X is the vector of p covariates, A is the treatment indication (say, $A = 1$ if treated, and $A = 0$ if untreated), T_d denotes the time to an event (say, death), and Y denotes the corresponding outcome by the time T_d. Y and T_d could be same. For example, if one is interested in survival analysis, one can take $Y = T_d$. However, if one is interested in life time event, such as life time health care cost, then Y is different from T_d. Let us denote C as censoring time. Let $T = \min\{T_d, C\}$ and $\delta = I\{T_d \leq C\}$ denote respectively the observed time-to-event or censoring, and censoring indication.

If $\delta = 1$, then T is the time-to-event and the outcome Y is observed. If $\delta = 0$, then T is the censoring time, and T_d and Y are unobserved and missing. Thus, when a subject is uncensored, we have observation on (X, A, T, Y, δ) with $\delta = 1$. However, when a subject is censored, we have observation on (X, A, T, δ) with $\delta = 0$, where Y is missing. We intend to estimate the ATE on the outcome Y. Suppose $Y^{(1)}$ be the potential outcome if the subject had received treatment, and $Y^{(0)}$ be the potential outcome if the subject had not received treatment. The ATE compares the average outcome if all subjects had received treatment (i.e., $A = 1$) with the average outcome if all subjects had not received the treatment (i.e., $A = 0$) . Mathematically, ATE can be written as (Hernán and Robins, 2020):

$$\mu = E(Y^{(1)}) - E(Y^{(0)}). \tag{4.1}$$

However, for each subject, we can at most observe one potential outcome, the one corresponding to treatment A the subject actually receives provided that the outcome is uncensored. That is, the observed Y is the potential outcome $Y^{(A)}$ given $\delta = 1$. Note that $E(Y^{(1)}) = E(Y^{(1)}|A = 1)Pr(A = 1) + E(Y^{(1)}|A = 0)Pr(A = 0)$, where the first term $E(Y^{(1)}|A = 1)$ can be further written as $E(Y^{(1)}|A = 1, \delta = 1)Pr(\delta = 1|A = 1) + E(Y^{(1)}|A = 1, \delta = 0)Pr(\delta = 0|A = 1)$, and $E(Y^{(1)}|A = 1, \delta = 1)$ can be directly estimated from observed data. However, all the other terms such as $E(Y^{(1)}|A = 1, \delta = 0)$ and $E(Y^{(1)}|A = 0)$ can not be directly calculated from observed data. To obtain these missing information, some conventional assumptions on missing data and causal inference are followed. Specifically, we assume that: (i) the outcome variable Y is missing at random, that is, the censoring C depends on the observed (X, A) but not the missing value Y itself; (ii) there is no unmeasured confounding variables, that is $(Y^{(0)}, Y^{(1)})$ is independent of A given X; (iii) positivity, that is, the subject has a non-zero probability of receiving each treatment given covariates X.

Under these assumptions, we develop a suitable method to estimate the ATE.

4.2.1 ATE estimates when there are censoring and confounding

In an observational study, we use the propensity score to balance the confounding variables. The term propensity score refers to the probability of receiving the treatment conditional on the baseline covariates X, that is,

$$p(X) = Pr(A = 1|X). \tag{4.2}$$

Note that we have observations on X and A for all subjects. The propensity score is estimated using the information only on X and A, thus we can obtain the propensity score estimation for each subject. If the entire study population are completely observed (i.e., no censoring), the IPTW method is applicable. The weight for subject i is defined as $w_i = \frac{A_i}{p(X_i)} + \frac{1-A_i}{1-p(X_i)}$. The ATE can be estimated by

$$\hat{\mu}_{(IPTW)} = \left(\sum_{i=1}^{N} \frac{A_i}{\hat{p}(X_i)}\right)^{-1} \sum_{i=1}^{N} \frac{A_i Y_i}{\hat{p}(X_i)} - \left(\sum_{i=1}^{N} \frac{1-A_i}{1-\hat{p}(X_i)}\right)^{-1} \sum_{i=1}^{N} \frac{(1-A_i)Y_i}{1-\hat{p}(X_i)}. \tag{4.3}$$

When the censoring is non-informative (i.e., the censoring time C is independent of the survival time T_d and outcome Y, and the subject is randomly censored), the ATE can still be estimated by IPTW method using the uncensored data. That is, ATE can be estimated as follows:

$$\hat{\mu}_{(IPTW)} = \left(\sum_{i=1}^{N} \frac{A_i \delta_i}{\hat{p}(X_i)}\right)^{-1} \sum_{i=1}^{N} \frac{A_i \delta_i Y_i}{\hat{p}(X_i)} - \left(\sum_{i=1}^{N} \frac{(1-A_i)\delta_i}{1-\hat{p}(X_i)}\right)^{-1} \sum_{i=1}^{N} \frac{(1-A_i)\delta_i Y_i}{1-\hat{p}(X_i)}. \tag{4.4}$$

When the censoring is informative (i.e., the censoring time C depends on the treatment A and the baseline covariates X), we should adjust for the informative censoring by weighting each uncensored subject with the inverse of the probability of uncensoring. That is, the weight is $\frac{\delta_i}{h(T_i;X_i,A_i)}$. Here $h(t;X,A) = Pr(C \geq t|X,A)$,

the probability of remaining uncensored at time t, which could be estimated using semi-parametric method such as Cox proportional model. The weight $\frac{\delta_i}{h(T_i;X_i,A_i)}$ makes the subject i (uncensored) represents $\frac{1}{\hat{h}(T_i:X_i,A_i)}$ subjects in the original study population, thus making the weighted uncensored sample similar to the original observed study population in baseline characteristics. The propensity score weighting $\left(\frac{A_i}{\hat{p}(X_i)} + \frac{1-A_i}{1-\hat{p}(X_i)}\right)$ balances the baseline characteristics between treatment group and control group. The double weights $\left(\frac{A_i}{\hat{p}(X_i)} + \frac{1-A_i}{1-\hat{p}(X_i)}\right) \times \frac{\delta_i}{\hat{h}(T_i;X_i,A_i)}$ adjust for both confounding and censoring under the identifiability conditions for (A, δ) conditional on X, that is, $Y^{(a,1)} \perp (A, \delta)|X$, joint positivity for $(A = a, \delta = 1)$ and consistency, for $a = 0, 1$ (Hernán and Robins, 2020). Here $Y^{(a,1)}$ denotes the potential outcome under treatment a and remaining uncensored. The ATE can be estimated by

$$\begin{aligned}\hat{\mu}_{(DW)} = &\left(\sum_{i=1}^{N} \frac{A_i\delta_i}{\hat{p}(X_i)\hat{h}(T_i;X_i,A_i)}\right)^{-1} \sum_{i=1}^{N} \frac{A_i\delta_i Y_i}{\hat{p}(X_i)\hat{h}(T_i;X_i,A_i))} \\ &- \left(\sum_{i=1}^{N} \frac{(1-A_i)\delta_i}{(1-\hat{p}(X_i))\hat{h}(T_i;X_i,A_i))}\right)^{-1} \sum_{i=1}^{N} \frac{(1-A_i)\delta_i Y_i}{(1-\hat{p}(X_i))\hat{h}(T_i;X_i,A_i))}.\end{aligned} \tag{4.5}$$

The ATE estimator based on the doubly weighting method is consistent if both propensity score and censoring model are correctly specified. However, if all the covariates are included in the propensity score model, the precision of the ATE estimates may suffer. In the following section 4.2.2, we present how to select the covariates for the propensity score estimation.

4.2.2 Variable selection for propensity score model

Suppose all the baseline covariates X can be classified into four categories: (1) the instrumental variables X_I , which are only related to the treatment selection; (2) the confounding factors X_C, which are related to both the treatment selection and outcome; (3) the prognostic variables X_P, which are only related to the outcome; and (4) the spurious variables X_S, which are neither related to treatment selection

nor to outcome. For these four different types of covariates, Brookhart et al. (2006) concluded that using the covariates related to the outcome model (i.e., X_C and X_P) in the propensity score model improves the performance of the propensity-score-based methods in estimating ATE (Brookhart et al., 2006). Thus, we propose to use the lasso method to select the covariates which are important to the survival time T_d, a precursor of Y, and then use the selected covariates to estimate the propensity scores. We propose using the followings steps:

(1). Construct the Cox proportional hazard model for the survival time T_d,

$$f(T_d|X, A) = f_0(T_d|X) \exp(A\alpha_0 + X'\alpha), \tag{4.6}$$

We use the lasso method to select the covariates by adding penalized regularization, which is implemented using the cv.glmnet() function in R package "glmnet". The selected covariates are denoted by $X_{(S_{out})}$.

(2). The propensity score is estimated from the following logistic regression model with the selected variables $X_{(S_{out})}$,

$$\text{logit}\{p(X)\} = X'_{(S_{out})}\beta, \tag{4.7}$$

$\hat{\beta}$ could be estimated by using the maximum likelihood method, and we can obtain the propensity score estimation:

$$\hat{p}_{(S_{out})}(X) = \frac{\exp(X'_{(S_{out})}\hat{\beta})}{1 + \exp(X'_{(S_{out})}\hat{\beta})}. \tag{4.8}$$

Note that in the Step (1), we can use the outcome model for Y instead of for T_d if they differ, and the censoring weights in the following section could be used to obtain a more accurate outcome model if the censoring is informative.

4.2.3 Estimate the probability of being uncensored

In the case that the censoring time C is related to the treatment A and baseline covariates X, we need to estimate the probability of remaining uncensored $h(T; X, A)$. Note that the "censoring" is the event we are interested in, so that the complete case indication is $1 - \delta$.

We propose to use the Cox proportional hazard model to estimate the probability of remaining uncensored $h(T; X, A)$:

$$g(C|X, A) = g_0(C|X) \exp(A\gamma_0 + X'\gamma). \tag{4.9}$$

Considering the high dimensional data, we consider the following procedures to select covariates for the censoring model:

M1: **Using covaraites selected from outcome model:** We investigate to use the selected covariates $X_{(S_{out})}$ from the outcome model to estimate the censoring probability using model (4.9) and the dataset $\{(X_{(S_{out},i)}, A_i, T_i, 1 - \delta_i), i = 1, 2, ..., N\}$. Let denote the resulting estimates as $\{\hat{h}(T_i; X_i, A_i), i = 1, 2, ..., N\}$;

M2: **Censoring based covariates:** We include all covariates in the model (4.9), however, we use the lasso method to select important covariates (say, $X_{(S_{cen})}$) and estimate $\{\hat{h}(T_i; X_i, A_i), i = 1, 2, ..., N\}$.

As an alternative, one may treat censoring as a binary variable, and estimate the probability of being uncensored: $\Pr(\delta = 1|A, X)$. Since the event $\{C \geq T_d\}$ is equivalent to $\{\delta = 1\}$, we can treat the censoring indication δ as a binary variable and build the following logistic regression model,

$$\text{logit}\{\Pr(\delta = 1|A, X)\} = A\theta_0 + X'\theta. \tag{4.10}$$

The selected covariates $X_{(S_{out})}$ based on the outcome model (4.6) could be used to

estimate $\Pr(\delta = 1|A, X)$. Likewise, we could also use the lasso method to estimate $\Pr(\delta = 1|A, X)$. However, when censoring time is available, the censoring probability is more likely dependent on time. Thus the time-dependent censoring model is more suitable than the binary model. Our simulation results (unshown) indeed show that time-dependent censoring model performs better than time-independent censoring model (4.10).

With the propensity scores estimated using the selected variables in Section 4.2.2 (say, $\hat{p}_{(S_{out})}(X)$) and the probability of being uncensored in this subsection (say $\hat{h}(T_i; X_i, A_i)$), we can estimate the ATE using the following expression:

$$\begin{aligned}\hat{\mu}_{(DW)}^{(S_{out})} = &\left(\sum_{i=1}^{N} \frac{A_i\delta_i}{\hat{p}_{(S_{out})}(X_i)\hat{h}(T_i; X_i, A_i)}\right)^{-1} \sum_{i=1}^{N} \frac{A_i\delta_i Y_i}{\hat{p}_{(S_{out})}(X_i)\hat{h}(T_i; X_i, A_i)} \\ &- \left(\sum_{i=1}^{N} \frac{(1-A_i)\delta_i}{(1-\hat{p}_{(S_{out})}(X_i))\hat{h}(T_i; X_i, A_i)}\right)^{-1} \sum_{i=1}^{N} \frac{(1-A_i)\delta_i Y_i}{(1-\hat{p}_{(S_{out})}(X_i))\hat{h}(T_i; X_i, A_i)}.\end{aligned} \tag{4.11}$$

4.3 Simulation study

In this simulation study, we evaluate the performance of different methods in Section 4.2, including the IPTW, doubly weighting (DW) methods with different propensity score and probability of remaining uncensored estimations. We consider two simulation scenarios: (1) non-informative censoring and (2) informative censoring. Under each scenario, we use 100 and 500 baseline covariates, separately, to examine the performance of the proposed methods when the number of covariates varies.

4.3.1 Simulation setting

In this simulation study, we simulate three instrument variables (X_I), four confounding factors (X_C) and three prognostic variables (X_P) from normal distribution with

mean zero and standard deviations at 0.25, 0.1, and 0.25, respectively. Then we generate $p-10$ (here p=100 or 500) spurious variables (X_S) from normal distribution with mean zero and standard deviations at 0.25. That is, the baseline covariates are denoted as $X = (X_I, X_C, X_P, X_S)'$. The treatment assignments A are generated from the logistic regression model which are related to the instrument variables (X_I) and confounding factors (X_C),

$$\text{logit}\{\Pr(A = 1|X)\} = X_I'\beta_I + X_C'\beta_C. \tag{4.12}$$

Here, $\beta_I = (0.65, 0.65, 0.65)'$ and $\beta_C = (0.4, 0.4, 0.40, 0.40)'$. The survival times T_d are generated from the weibull distribution based on the confounding factors (X_C) and the prognostic variables (X_P) from the following model:

$$T_d = \left(\frac{-log(u)}{\lambda \exp(A\alpha_0 + X_C'\alpha_C + X_P'\alpha_P)}\right)^{\frac{1}{\eta}}. \tag{4.13}$$

Here $u \sim U(0,1)$, $\lambda = 0.002$, $\eta = 2$, $\alpha_C = (1,1,1,1)$ and $\alpha_P = (1.3, 1.3, 1.3)$ (Bender et al., 2005). The α_0 captures the treatment effect between the treatment $A = 1$ and $A = 0$. In the simulation study, α_0 is set to be 0, which indicates that the true ATE is 0. The sample size is set to be 1000. In this simulation setting, we assume the corresponding outcome Y equals to survival time T_d.

For the censoring mechanism, we consider two scenarios based on their relationship with the treatment A and baseline covariates X. The probability of censoring is set approximately 20%.

- **Scenario I.** Non-informative censoring (i.e., Figure 4.1): the censoring times C are simulated from the weibull distribution with the formula $C = \left(\frac{-log(u)}{\lambda_{cens}}\right)^{\frac{1}{\eta_{cens}}}$, where $u \sim U(0,1)$, $\lambda_{cens} = 0.04$, $\eta_{cens} = 0.6$. The censoring probability is independent of X and A.

- **Scenario II.** Informative censoring (i.e., Figure 4.2): the censoring times C are simulated from the weibull distribution with the following expression:

$$C = \left(\frac{-log(u)}{\lambda_{cens} \exp(A\gamma_0 + X_C'\gamma_C + X_P'\gamma_P)} \right)^{\frac{1}{\eta_{cens}}}. \tag{4.14}$$

Here $u \sim U(0,1)$, $\lambda_{cens} = 0.03$, $\eta_{cens} = 0.5$, $\gamma_C = (-2.0, -2.0, -2.0, -2.0)$ and $\gamma_P = (0.5, 0.5, 0.5)$.

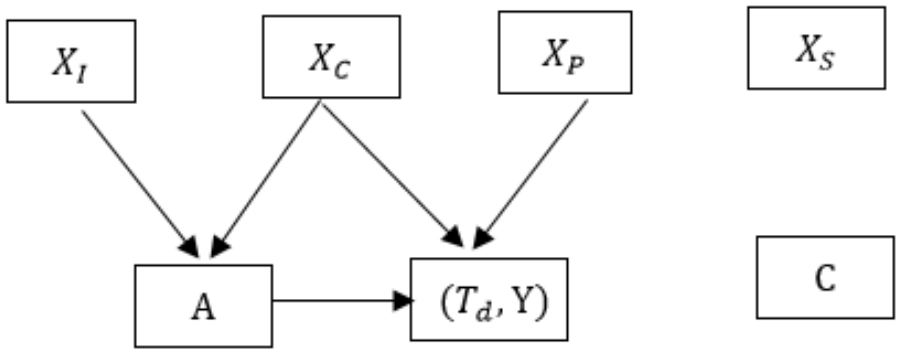

Figure 4.1: Scenario I: Non-informative censoring

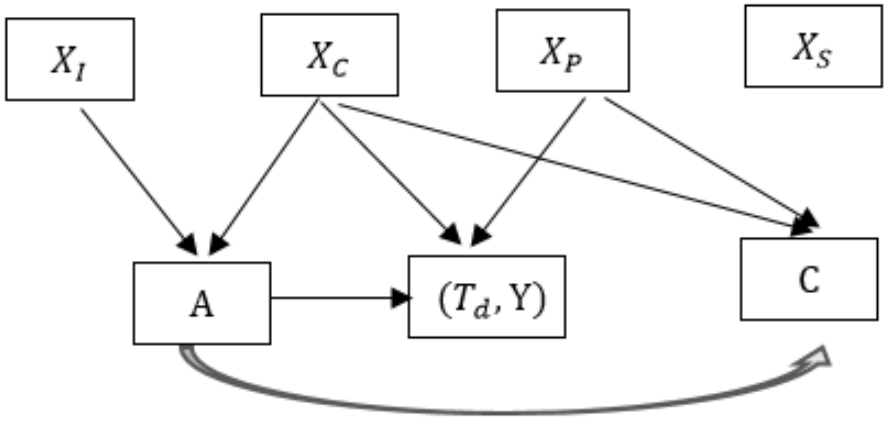

Figure 4.2: Scenario II: Informative censoring

Based on the above simulation settings, the simulation algorithm to evaluate the performances of ATE estimations are described as follows:

Step 1. Generate the vectors of 500 covariates X_i, for $i = 1, 2, ..., N$. However, we only use the first 10 covariates to generate the treatment assignment A_i via the logistic model (4.12) and generate the survival time T_{di} via the Cox proportional model (4.13), for $i = 1, 2, ..., 1000$.

Step 2. Generate the non-informative censoring times C from Scenario I or the informative censoring C from Scenario II. From the survival times generated in Step 1 and the censoring times generated in this Step, the time-to-event or censoring time $T_i = \min\{T_{di}, C_i\}$, and the censoring indication $\delta_i = I\{C_i \geq T_{di}\}$ for $i = 1, 2, ..., 1000$.

Step 3. Estimate the propensity scores using the following covarites: (i) the confounding factors X_C only; (ii) the variables (X_C, X_P) only; (iii) the first 100 covariates; (iv) all 500 covariates; (v) the selected covariates based on the outcome model from the first 100 covariates; and (vi) the selected covariates based on the outcome model from all 500 covariates. The six PS estimators are denoted by $PS_{(Conf)}$, $PS_{(Out)}$, $PS_{(All100)}$, $PS_{(All500)}$, $PS_{(S_{out}100)}$, and $PS_{(S_{out}500)}$, respectively.

Step 4. Estimate the probability of remaining uncensored $h(T; X, A)$ using the Cox proportional hazard models but with the following nine specifications of covariates in the model:

(1). no covariates;

(2). confounding factors X_C only;

(3). confounding factors X_C and prognostic variables X_P only;

(4). the first 100 covariates X_{All100};

(5). all 500 covariates X_{All500};

(6). covariates selected based on the outcome-model using the first 100 covariates $X_{(S_{out}100)}$;

(7). covariates selected based on the outcome-model using all 500 covariates $X_{(S_{out}500)}$;

(8). covariates selected based on the censoring-model using the first 100 covariates $X_{(S_{cen}100)}$;

(9). covariates selected based on censoring-model using all 500 covariates $X_{(S_{cen}500)}$.

Step 5. Estimate the probability of remaining uncensored $\Pr(\delta = 1|X, A)$ using the logistic regression model but with the same types of covariates as in Step 4. Thus, we have nine sets of $\Pr(\delta = 1|X, A)$ estimators for each subject.

Step 6. Estimate ATE using the IPTW method (4.4) without accounting for the censoring, only use the observed data with $\delta = 1$.

Step 7. Estimate ATE using the doubly weighting method (4.5) with different propensity scores estimates and different probabilities of uncensoring in Steps 3-5.

For each scenario (Scenario I and II), we repeat the simulation steps 1-7 1000 times. We report the bias and standard error of the ATE estimates from different methods.

4.3.2 Simulation results

We first present the simulation results for Scenario II (i.e., informative censoring) in Figures 4.3 and 4.4, and Table 4.1.

The boxplots of 1000 ATE estimates under Scenario II are shown in Figure 4.3, where panels A and B correspond to the number of covariates $p = 100$ and $p = 500$, respectively. The legend "PS" explains three different types of propensity scores: (i) the true propensity score ("True") in the data generating process in Step

1, Section 4.3.1; (ii) all the covariates ("*All X*") are included in the propensity estimation model; (iii) only the selected variables based on outcome model ("S_{out} *X*") are included in the propensity estimation model. Each panel in Figure 4.3 shows the simulation results based on seven different ATE estimation methods. The first block $IPTW{:}True$ represents the IPTW method using (4.4) and the true propensity score. The second block"$IPTW{:}Est$" refers to the IPTW method with two different estimates of propensity scores: "*All X*" and "S_{out} *X*". The third one shows the results of double weights (DW) method in formula (4.11) but with true propensity score and true probability of being uncensored. The remaining four blocks show the results from DW methods but with different methods to estimate the probability of uncensoring: (1)$DW{:}PC_{X_0}$-no covariate in the probability of uncensoring estimation model (4.9); (2)$DW{:}PC_{X_{ps}}$-same covariates as that of the propensity score estimation model; (3)$DW{:}PC_{X_{S_{cen}}}$-covariates selected from censoring model (M2, section 4.2.3); (4)$DW{:}PC_{X_{S_{out}}}$-covariates selected from outcome model (M1, section 4.2.3). Figure 4.4 shows the the mean of the 1000 estimates from 1000 simulated dataset and their 95% confidence interval for each ATE estimation method under Scenario II.

Based on the simulation results shown in Figures 4.3-4.4 and Table 4.1, we draw the following conclusions:

(1). By comparing IPTW and DW methods with true propensity score and true probability of uncensoring (i.e., $IPTW{:}True$ versus $DW{:}True$), the DW method provides a smaller bias with similar standard deviation, which are shown in the first and third blocks of panels A and B in Figure 4.3 and Table 4.1.

(2). If the censoring is informative, when the number of covariates becomes larger (p=100 versus p=500), the variable selection for the propensity score model clearly reduces the bias and the variance of the ATE estimates, comparing to those which include all the covaraites into the propensity score estimation model.

(3). If the propensity score is estimated based on the selected covariates, by comparing all the blue boxplots "S_{out} X" in Figure 4.3, the results based on $IPTW$:Est method are still biased, however, the performance of four DW methods (DW:PC_{X_0}, DW:$PC_{X_{ps}}$, DW:$PC_{X_{S_{cen}}}$, DW:$PC_{X_{S_{out}}}$) depend on the probability of uncensoring estimation method. The performance of DW:PC_{X_0} methods doesn't perform well, compare to other three DW methods. Because, the probability of uncensoring estimation model in DW:PC_{X_0} is not correctly specified when no-covariate is included.

Based on the results for the informative censoring, we conclude that it is necessary to use the censoring weight to account for the censoring information. When the number of covariates becomes large, it is beneficial to select the outcome-related covariates to the propensity score estimation model. The simulation results for Scenario I (i.e., non-informative censoring) are reported in Figures A1.8-A1.9 and Table A1.1. It is obviously that when the censoring is not informative, all the methods provide unbiased estimates. We also note that the methods with variable selection provide more accurate estimates when the number of covariates is large.

Based on the simulation results (not shown), if we use logistic regression model to estimate the probability of uncensoring (i.e., $\Pr(\delta = 1|A, X)$ in Step 5 in Section 4.3.1), the results based on all methods are biased. The reason is that the true censoring time C is generated from the Cox proportionals model, and the complete case indication δ is related to the survival time T_d. The logistic regression model is not sufficient to capture the time-dependent censoring. In the logistic regression model, we only use the data (X, A, δ) without using censoring time. To make valid estimation for ATE using the doubly weighting method, both the propensity score model and probability of uncensoring model must be correctly specified.

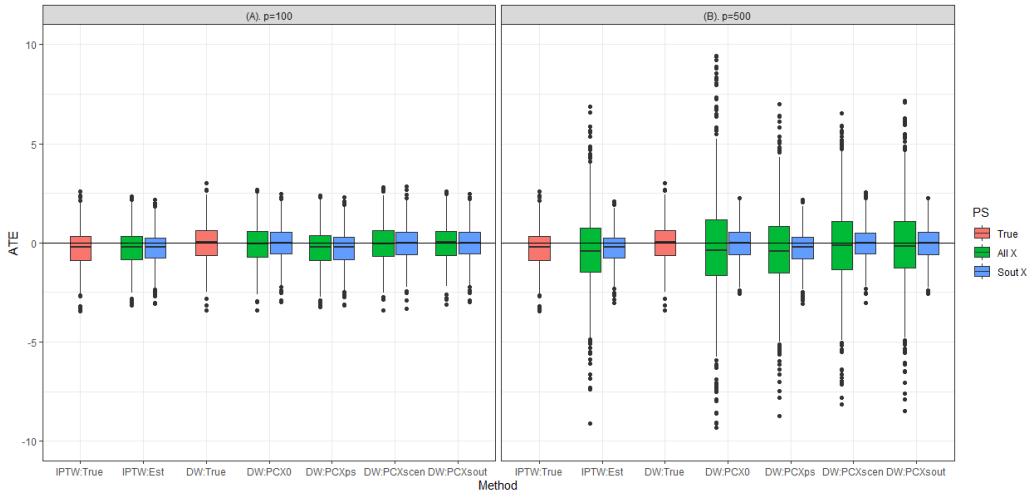

Figure 4.3: The boxplots of 1000 ATE estimates based on IPTW and DW methods, combination with different sets of covariates in the propensity score model, and different sets of covariates in the probability of uncensoring model, under Scenario II.

4.4 Case study

In case study, we are interested in comparing the treatment effect on survival time between surgery and chemotherapy treatment for pancreas patients. We use the SEER-Medicare data to create a cohort of pancreas patients whose diagnosis dates were between Jan.1, 2006 and Dec. 31, 2013. The SEER-Medicare data are the linkage of two large population-based sources of data, that is, SEER cancer registry and Medicare, which provide detailed information about Medicare beneficiaries with cancer. The SEER cancer registries denote the Surveillance, Epidemiology and End Results of cancer registries, which collect clinical, demographic and cause of death information for persons with cancer and the Medicare claims for covered health care services from the time of a person's Medicare eligibility until death. For persons reported to a SEER registry who were aged 65 or older, 94% have been linked to Medicare and their Medicare claims have been extracted (Warren et al., 2002).

In this study, we obtain the cohort of pancreas patients from the patient entitlement and diagnosis summary file (PEDSF, SEER base file), with the inclu-

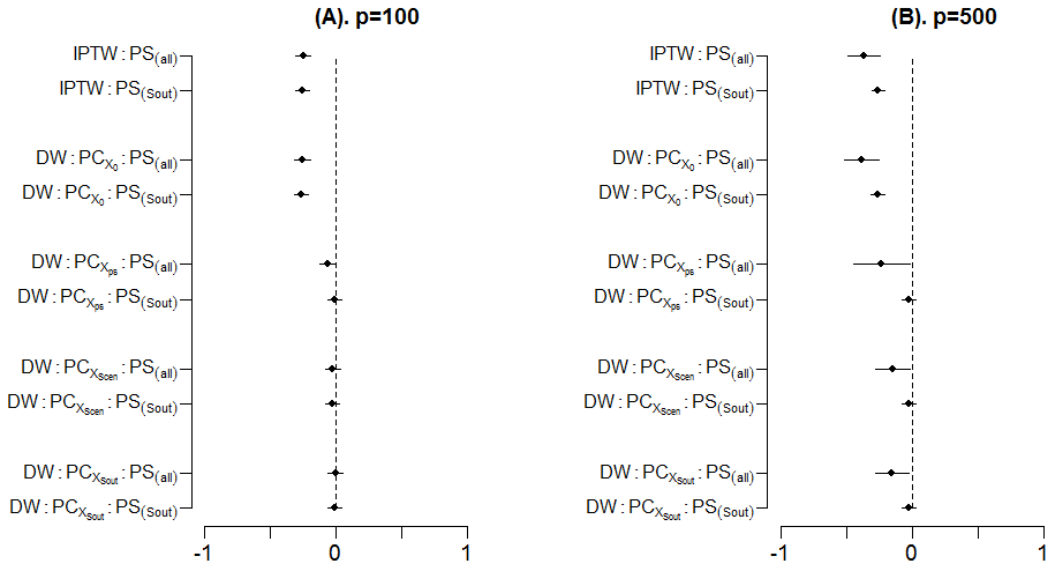

Figure 4.4: ATE estimates and their 95% CI of ATE estimates for p=100 and 500 under Scenario II.

sion criterion that each patient should have histological code 8140 (adenocarcinoma, NOS), 8141 (scirrhous adenocarcinoma), 8143 (superficial spreading adenocarcinoma) or 8147 (basal cell adenocarcinoma), and behavioural code 3 (malignant, primary). Based on these inclusion criterion, there are 3745 pancreas patients with complete information on their baseline covariates: gender, race, state, urbrur, age. Among them, there are 3137 patients whose survival times are completely observed. We calculate the Charlson comorbidity score for these patients from the claims data: inpatient hospitalizations (MEDPAR), outpatient facilities, physician claims.

The ICD-9 codes for pancreas surgery are pancreatotomy (52.01, 52.09), partial pancreatotomy (52.51, 52.52, 52.53, 52.59) and total pancreatotomy (52.6). The chemotherapy codes are listed as followings: 5201, 5209, 5251, 5252, 5253, 5259, 526,J9201, J9264, J8999, J8520, J8521, J9060, J9228, J9271, J9299, J9035, J9045, J9280, J9055, J000, J9311, J8530, J9040, J8560, J9178, J9200, J9312, J9070, J9310, J9181. All patients are classified into four treatment groups based on their treatment received: (1) SURG: only receive surgery treatment; (2) CHEM: only receive the chemotherapy; (3) SURG/CHEM: first receive surgery then chemotherapy treatment;

(4) CHEM/SURG: first receive chemotherapy then surgery treatment. In this study, we only estimate the average treatment effect between the "CHEM" and "SURG" groups.

We summarize each continuous variable by calculating mean and standard error; and we summarize each categorical variable by counts and percentages, stratified by the treatment groups. The summary information is shown under "Observed sample" in Table 4.2 and Table 4.3.

Before estimating the treatment effect, we need to estimate the propensity score and the probability of being uncensored. The propensity score is estimated from two variable specifications: (1) including all the covariates in the model, denoted by $PS_{(All)}$; and (2) including the covariates related to the outcomes, denoted by $PS_{(S_{out})}$. For the probability of being uncensored, we also utilize two variable specifications: (1) including the outcome-related covariates, denoted by $PC_{X_{S_{cen}}}$; and (2) including the censoring-related covariates, denoted by $PC_{X_{S_{out}}}$. Then we use the doubly weighting method in formula (4.11) to compare the mean survival time between "SURG" and "CHEM" groups, which are shown in the first two columns of Table 4.4. In addition, we generate 100 spurious covariates from the standard normal distribution and combine these with the baseline variables obtained from the SEER-Medicare dataset. The motivation is to examine the performance of variable selection method. Given the new "baseline covariates", we repeat the above doubly weighting method and variable selection method to estimate the difference between mean survival time of the two treatments. The estimations are shown under the last two columns of Table 4.4.

Based on Table 4.4, it is clear that when the $\boldsymbol{DW}$:$PC_{X_{S_{cen}}}$ coupled with $PS_{(S_{out})}$ is used, the ATE estimates are similar no matter whether we use the 100 generated spurious covariates or not. However, when the variable selection is not applied to propensity score estimation, the ATE estimate depends on whether we use

the spurious covariates or not. Based on our simulation study, the variable selection for propensity score estimation does improve the accuracy of ATE estimation. We make the statistic inference for ATE estimate based on the method with variable selection. We conclude that the surgery treatment helps patients to survive longer, and the difference of mean survival time is around 188 days.

4.5 Discussion

In this study, we propose the doubly weighting method coupled with the variable selection method to estimate ATE in the context of survival analysis, when there are confounding and censoring. To estimate the propensity scores, we use the lasso method to select the covariates relevant to the outcomes as the variables in the propensity score estimation model. We also investigate two different methods to select the covariates for the probability of uncensoring estimation model. Based on the simulation results, we concluded that the probability of uncensoring estimation model should be correctly specified in order to capture the informative censoring sufficiently. We also clearly illustrate the importance of variable selection for the propensity score model, particularly when the number of covariates is large. If we include all the covariates in the propensity score estimation model, the ATE estimates are biased and their standard errors are larger comparing to the case when only selected variables are included in the propensity score model. The doubly weighting method performs much better when we use the selected covariates in the propensity score and probability of uncensoring estimation models than we do not make variable selection.

We will expand the research from the following two aspects. First, we will explore the doubly robust method coupled with the variable selection method. Second, in the current simulation setting, we only explore the simple situation for outcome Y, that is, $Y = T_d$. In future, we will design the simulation to consider the more complex outcome, say, the medical cost since the diagnosis of a disease.

Table 4.1: Bias and standard error (S.E.) of ATE estimates based on IPTW and DW methods under Scenario II: informative censoring.

	p=100		p=500	
	Bias	S.E.	Bias	S.E.
True ATE	0.0001	0.015		
$\boldsymbol{IPTW}$:$True$	-0.2699	0.0287		
$\boldsymbol{IPTW}$:Est				
$PS_{(Conf)}$	-0.2724	0.0285		
$PS_{(Out)}$	-0.2493	0.0255		
$PS_{(All)}$	-0.2482	0.0279	-0.3691	0.0633
$PS_{(S_{out})}$	-0.2549	0.0252	-0.2589	0.0249
$\boldsymbol{DW}$:$True$	-0.0096	0.0297		
$\boldsymbol{DW}$:PC_{X_0}				
$PS_{(Conf)}$	-0.2784	0.0298		
$PS_{(Out)}$	-0.2546	0.0267		
$PS_{(All)}$	-0.2535	0.0294	-0.3865	0.0662
$PS_{(S_{out})}$	-0.2603	0.0265	-0.2638	0.0262
$\boldsymbol{DW}$:$PC_{X_{ps}}$				
$PS_{(Conf)}$	0.1171	0.0304		
$PS_{(Out)}$	0.0014	0.0264		
$PS_{(All)}$	-0.0602	0.0298	-0.2335	0.1088
$PS_{(S_{out})}$	-0.0103	0.0261	-0.026	0.0261
$\boldsymbol{DW}$:$PC_{X_{S_{cen}}}$				
$PS_{(Conf)}$	-0.0389	0.0301	-0.0394	0.0301
$PS_{(Out)}$	-0.0154	0.0272	-0.0159	0.0272
$PS_{(All)}$	-0.0234	0.0297	-0.148	0.0661
$PS_{(S_{out})}$	-0.0242	0.0269	-0.0291	0.0266
$\boldsymbol{DW}$: $PC_{X_{S_{out}}}$				
$PS_{(Conf)}$	-0.0279	0.0293	-0.0399	0.0296
$PS_{(Out)}$	-0.0043	0.0264	-0.0166	0.0267
$PS_{(All)}$	-0.0033	0.029	-0.1541	0.0652
$PS_{(S_{out})}$	-0.0103	0.0261	-0.026	0.0261

Note: the blank space under the column "p=500" indicates the exactly same result as those under the column "p=100".

Table 4.2: The summary of the variables under the two treatment groups in the observed sample and pseudo population.

		Observed sample		Pseudo population	
		CHEM	SURG	CHEM	SURG
Gender	Male	551	945	1824.6	1602.3
		(47.9%)	(47.6%)	(48.9%)	(47.7%)
	Female	599	1042	1906	1756.6
		(52.1%)	(52.4%)	(51.1%)	(52.3%)
Race	White	968	1734	3179.6	2928.7
		(84.2%)	(87.3%)	(85.2%)	(87.2%)
	Others	182	253	551.5	430.2
		(15.8%)	(12.7%)	(14.8%)	(12.8%)
Urbrur	Metro	1013	1685	3228.3	2901.8
		(88.1%)	(84.8%)	(86.5%)	(86.4%)
	Urban	119	263	432.2	399.4
		(10.3%)	(13.2%)	(11.6%)	(11.9%)
	Rural	18	39	70.1	57.7
		(1.6%)	(2%)	(1.9%)	(1.7%)
Age		72.5	73.1	72.7	73
		0.2	0.2	0.1	0.1
NCI		1.3	1.4	72.7	73
		0.04	0.03	0.1	0.1

Table 4.3: The summary of comorbidity scores under the two treatment groups in the observed sample and pseudo population.

	Observed sample				Pseudo population			
	CHEM		SURG		CHEM		SURG	
	Sample	(%)	Sample	(%)	Sample	(%)	Sample	(%)
ACUTE_MI:0	1147	99.7	1975	99.4	3709.5	99.4	3343.3	99.5
ACUTE_MI:1	3	0.3	12	0.6	21.1	0.6	15.6	0.5
AIDS:0	1146	99.7	1983	99.8	3722.6	99.8	3351	99.8
AIDS:1	4	0.3	4	0.2	8	0.2	7.9	0.2
CHF:0	1067	92.8	1841	92.7	3474.3	93.1	3121.1	92.9
CHF:1	83	7.2	146	7.3	256.3	6.9	237.7	7.1
COPD:0	965	83.9	1654	83.2	3150.5	84.4	2792.1	83.1
COPD:1	185	16.1	333	16.8	580.2	15.6	566.8	16.9
CVD:0	1093	95	1880	94.6	3551.2	95.2	3191.3	95
CVD:1	57	5	107	5.4	179.5	4.8	167.6	5
DEMENTIA:0	1141	99.3	1969	99.1	3703.2	99.3	3332.3	99.2
DEMENTIA:1	9	0.8	18	0.9	27.4	0.7	26.6	0.8
DIABETES:0	694	60.3	1280	64.4	2268.5	60.8	2184.3	65
DIABETES:1	456	39.7	707	35.6	1462.2	39.2	1174.6	35
DIABETES_COMP:0	1040	90.4	1823	91.7	3402.2	91.2	3085.8	91.9
DIABETES_COMP:1	110	9.6	164	8.3	328.4	8.8	273.1	8.1
HISTORY_MI:0	1118	97.2	1916	96.4	3615.4	96.9	3245.2	96.6
HISTORY_MI:1	32	2.8	71	3.6	115.3	3.1	113.7	3.4
LIVER_DISEASE:0	1144	99.5	1982	99.7	3715.1	99.6	3349.2	99.7
LIVER_DISEASE:1	6	0.5	5	0.3	15.6	0.4	9.7	0.3
MILD_LIVER_DISEASE:0	1134	98.6	1967	99	3689.5	98.9	3334	99
MILD_LIVER_DISEASE:1	16	1.4	20	1	41.2	1.1	34.9	1
PARALYSIS:0	1149	99.9	1976	99.4	3716.4	99.6	3346.3	99.6
PARALYSIS:1	1	0.1	11	0.6	14.2	0.4	12.6	0.4
PVD:0	1062	92.3	1771	89.1	3408	91.4	3034.9	90.4
PVD:1	88	7.7	216	10.9	322.7	8.6	324	9.6
RENAL_DISEASE:0	1079	93.8	1844	92.8	3526.1	94.5	3127.7	93.1
RENAL_DISEASE:1	71	6.2	143	7.2	204.5	5.5	231.2	6.9
RHEUM_DISEASE:0	1123	97.7	1933	97.3	3645.9	97.7	3273.6	97.5
RHEUM_DISEASE:1	27	2.3	54	2.7	84.8	2.3	85.3	2.5
ULCERS:0	1130	98.3	1944	97.8	3669.6	98.4	3289.9	97.9
ULCERS:1	20	1.7	43	2.2	61	1.6	69	2.1

Table 4.4: ATE estimates in the case study

	Baseline variables		+100 spurious variables	
	ATE	Standard Error	ATE	Standard Error
$\boldsymbol{DW}$:$PC_{X_{S_{out}}}$				
$PS_{(All)}$	184.0	19.3	191.2	19.7
$PS_{(S_{out})}$	183.6	19.3	188.0	19.6
$\boldsymbol{DW}$:$PC_{X_{S_{cen}}}$				
$PS_{(All)}$	182.7	19.8	193.5	19.6
$PS_{(S_{out})}$	**188.7**	19.6	**187.1**	19.6

REFERENCES

Abdia, Y. (2016). Propensity score based methods for estimating the treatment effects based on observational studies.

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., and Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5):967–985.

Agresti, A. (2012). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley Sons, Inc.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based g-computation. *Multivariate behavioral research*, 47(1):115–135.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849.

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258.

Austin, P. C. and Schuster, T. (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical methods in medical research*, 25(5):2214–2237.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to

simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury.

CDC (2013). *The BRFSS Data User Guide*. Atlanta, GA: Department of Health and Human Services.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Datta, S., Pihur, V., and Datta, S. (2010). An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data. *BMC bioinformatics*, 11(1):427.

Dauchet, L., Amouyel, P., and Dallongeville, J. (2009). Fruits, vegetables and coronary heart disease. *Nature Reviews Cardiology*, 6(9):599–608.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Dietary Guidelines Advisory Committee (2015). *Dietary Guidelines for Americans 2015-2020*. Government Printing Office.

Ertefaie, A., Asgharian, M., and Stephens, D. A. (2018). Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, 6(1).

Franklin, J. M., Eddings, W., Glynn, R. J., and Schneeweiss, S. (2015). Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *American journal of epidemiology*, 182(7):651–659.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, L. M., Furberg, C., DeMets, D. L., Reboussin, D. M., Granger, C. B., et al. (2010). *Fundamentals of clinical trials*, volume 4. Springer.

Giannoudis, P. V., Dinopoulos, H., and Tsiridis, E. (2005). Bone substitutes: an update. *Injury*, 36(3):S20–S27.

Gibson, S., McLeod, I., Wardlaw, D., and Urbaniak, S. (2002). Allograft versus autograft in instrumented posterolateral lumbar spinal fusion: a randomized control trial. *Spine*, 27(15):1599–1603.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, 29(1):205–220.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton: Chapman Hall/CRC.

Horwitz, R. I. (1987). The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. *Journal of chronic diseases*, 40(1):91–99.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

Li, J., Handorf, E., Bekelman, J., and Mitra, N. (2016). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in medicine*, 35(12):1985–1999.

Lin, R. S. and León, L. F. (2017). Estimation of treatment effects in weighted log-rank tests. *Contemporary clinical trials communications*, 8:147–155.

Lumley, T. et al. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Moore, L. V., Dodd, K. W., Thompson, F. E., Grimm, K. A., Kim, S. A., and Scanlon, K. S. (2015). Using behavioral risk factor surveillance system data to estimate the percentage of the population meeting us department of agriculture food patterns fruit and vegetable intake recommendations. *American Journal of Epidemiology*, 181(12):979–988.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., and Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety*, 20(6):551–559.

Pihur, V., Datta, S., and Datta, S. (2009). Rankaggreg, an r package for weighted rank aggregation. *BMC bioinformatics*, 10(1):62.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, volume 24, page 3. San Francisco CA.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29(3):343–367.

Schaubel, D. E. and Wei, G. (2011). Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics*, 67(1):29–38.

Setodji, C. M., McCaffrey, D. F., Burgette, L. F., Almirall, D., and Griffin, B. A. (2017). The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology (Cambridge, Mass.)*, 28(6):802–811.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.

Shah, J., Datta, S., and Datta, S. (2014). A multi-loss super regression learner (msrl) with application to survival prediction using proteomics. *Computational Statistics*, 29(6):1749–1767.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Wang, C., Dominici, F., Parmigiani, G., and Zigler, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3):654–665.

Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B., and Riley, G. F. (2002). Overview of the seer-medicare data: content, research applications, and generalizability to the united states elderly population. *Medical care*, pages IV3–IV18.

WHO (2003). *Diet, Nutrition, and the Prevention of Chronic Diseases: Report of a Joint WHO/FAO Expert Consultation*, volume 916. World Health Organization.

Xie, J. and Liu, C. (2005). Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in medicine*, 24(20):3089–3110.

Xu, S., Shetterly, S., Powers, D., Raebel, M. A., Tsai, T. T., Ho, P. M., and Magid, D. (2012). Extension of kaplan-meier methods in observational studies with time-varying treatment. *Value in Health*, 15(1):167–174.

Yan, X., Abdia, Y., Datta, S., Kulasekera, K., Ugiliweneza, B., Boakye, M., and Kong, M. (2019). Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in Medicine*, 38:2828–2846.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation:

Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

APPENDIX

Appendix 1

This section includes the additional figures and tables in Chapters 2-4.

(τ_1, τ_2)		(0, 0)									(0, 0.5)								
Comparison groups		2 vs 1			3 vs 1			3 vs 2			2 vs 1			3 vs 1			3 vs 2		
True ATE		0			0			0			0			0.5			0.5		
		EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE
IPW	Mul	-0.001	0.043	0.075	-0.001	0.042	0.071	0	0.033	0.065	0.001	0.039	0.074	0.502	0.037	0.07	0.5	0.03	0.064
	RF	0.035	0.037	0.066	0.044	0.032	0.061	0.009	0.032	0.061	0.035	0.036	0.065	0.544	0.033	0.061	0.509	0.034	0.061
	GBM	0.038	0.033	0.065	0.037	0.03	0.06	-0.001	0.028	0.06	0.039	0.033	0.065	0.538	0.031	0.06	0.499	0.029	0.06
	CBPS	0.035	0.03	0.068	0.03	0.028	0.064	-0.005	0.025	0.062	0.035	0.03	0.068	0.531	0.029	0.064	0.496	0.025	0.061
	MinMean	0.017	0.034	0.071	0.015	0.033	0.066	-0.001	0.026	0.063	0.016	0.035	0.071	0.516	0.034	0.066	0.5	0.027	0.063
	MinMax	0.015	0.035	0.071	0.014	0.034	0.067	-0.001	0.027	0.063	0.015	0.035	0.071	0.515	0.034	0.067	0.5	0.028	0.063
DR	Mul	0	0.01	0.01	0	0.01	0.009	0	0.009	0.009	0	0.011	0.01	0.5	0.01	0.009	0.5	0.009	0.009
	RF	0	0.01	0.009	0	0.009	0.008	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	GBM	0	0.01	0.007	0	0.009	0.006	0	0.009	0.006	0	0.01	0.007	0.5	0.009	0.006	0.5	0.009	0.006
	CBPS	0	0.01	0.009	0	0.009	0.008	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	MinMean	0	0.01	0.009	0	0.009	0.009	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.009	0.5	0.009	0.008
	MinMax	0	0.01	0.009	0	0.009	0.009	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.009	0.5	0.009	0.008
enDR	Mul	0	0.01	0.01	0	0.01	0.009	0	0.009	0.009	0	0.011	0.01	0.5	0.01	0.009	0.5	0.009	0.009
	RF	0	0.01	0.009	0	0.009	0.008	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	GBM	0	0.01	0.007	0	0.009	0.006	0	0.009	0.006	0	0.01	0.007	0.5	0.009	0.006	0.5	0.009	0.006
	CBPS	0	0.01	0.009	0	0.009	0.008	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	MinMean	0	0.01	0.009	0	0.009	0.009	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.009	0.5	0.009	0.008
	MinMax	0	0.01	0.009	0	0.009	0.009	0	0.009	0.008	0	0.01	0.009	0.5	0.009	0.009	0.5	0.009	0.008
enOM		0	0.01	0.032	0	0.009	0.029	0	0.008	0.03	0	0.01	0.032	0.5	0.009	0.029	0.5	0.009	0.031

Figure A1.1: Simulation results for Scenario AA (i.e., GPS_A and Out_A), where EST and SE are, respectively, the average of 1000 estimated ATEs and their standard errors. Emp.SE is the standard deviation of the 1000 estimated ATEs.

(τ_1, τ_2)		(0, 0)									(0, 0.5)								
Comparison groups		2 vs 1			3 vs 1			3 vs 2			2 vs 1			3 vs 1			3 vs 2		
True ATE		0			0			0			0			0.5			0.5		
		EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE
IPW	Mul	0.008	0.171	0.226	0.005	0.171	0.218	-0.004	0.157	0.205	0.004	0.173	0.225	0.5	0.168	0.218	0.496	0.152	0.206
	RF	-0.127	0.113	0.203	-0.074	0.104	0.19	0.053	0.111	0.193	-0.132	0.114	0.202	0.419	0.1	0.19	0.551	0.105	0.192
	GBM	-0.102	0.104	0.201	-0.055	0.103	0.189	0.046	0.098	0.191	-0.107	0.106	0.2	0.438	0.101	0.188	0.545	0.092	0.191
	CBPS	0.02	0.151	0.21	0.031	0.15	0.2	0.011	0.143	0.195	0.012	0.156	0.209	0.524	0.149	0.2	0.512	0.137	0.194
	MinMean	0.002	0.163	0.217	0.007	0.158	0.208	0.005	0.145	0.2	-0.005	0.167	0.216	0.502	0.156	0.208	0.507	0.141	0.2
	MinMax	-0.002	0.163	0.218	0.004	0.159	0.209	0.005	0.148	0.201	-0.005	0.165	0.217	0.499	0.153	0.209	0.505	0.142	0.201
DR	Mul	0	0.177	0.162	-0.002	0.172	0.156	-0.002	0.144	0.137	0.005	0.187	0.163	0.502	0.177	0.158	0.497	0.141	0.138
	RF	-0.016	0.113	0.134	-0.021	0.103	0.126	-0.006	0.097	0.128	-0.015	0.116	0.134	0.48	0.106	0.125	0.495	0.092	0.127
	GBM	-0.016	0.111	0.101	-0.018	0.104	0.098	-0.002	0.095	0.096	-0.014	0.114	0.1	0.483	0.106	0.098	0.498	0.092	0.096
	CBPS	-0.01	0.156	0.137	-0.008	0.15	0.131	0.002	0.137	0.125	-0.009	0.164	0.137	0.493	0.152	0.131	0.502	0.131	0.125
	MinMean	-0.005	0.163	0.148	-0.005	0.156	0.141	-0.001	0.137	0.129	-0.003	0.17	0.148	0.497	0.158	0.142	0.5	0.132	0.13
	MinMax	-0.006	0.16	0.148	-0.008	0.155	0.142	-0.002	0.137	0.131	-0.003	0.169	0.149	0.496	0.157	0.143	0.499	0.132	0.131
enDR	Mul	-0.022	0.078	0.046	-0.01	0.069	0.043	0.012	0.068	0.043	-0.02	0.077	0.047	0.491	0.07	0.044	0.511	0.068	0.044
	RF	-0.028	0.07	0.041	-0.014	0.061	0.037	0.013	0.062	0.039	-0.027	0.068	0.041	0.486	0.061	0.037	0.512	0.061	0.039
	GBM	-0.029	0.069	0.033	-0.014	0.061	0.031	0.015	0.062	0.032	-0.028	0.068	0.033	0.486	0.061	0.032	0.514	0.061	0.032
	CBPS	-0.024	0.074	0.041	-0.011	0.066	0.038	0.013	0.066	0.039	-0.023	0.073	0.041	0.489	0.066	0.039	0.512	0.065	0.04
	MinMean	-0.023	0.075	0.043	-0.011	0.067	0.04	0.013	0.067	0.041	-0.021	0.075	0.044	0.49	0.067	0.041	0.511	0.066	0.041
	MinMax	-0.023	0.075	0.043	-0.011	0.067	0.04	0.012	0.067	0.041	-0.021	0.074	0.044	0.49	0.067	0.041	0.511	0.066	0.041
enOM		-0.035	0.073	0.614	-0.015	0.064	0.603	0.019	0.066	0.532	-0.034	0.072	0.604	0.484	0.064	0.593	0.518	0.065	0.524

Figure A1.2: Simulation results for Scenario AB (i.e., GPS_A and Out_B), where EST and SE are, respectively, the average of 1000 estimated ATEs and their standard errors. Emp.SE is the standard deviation of the 1000 estimated ATEs.

(τ_1, τ_2)		(0, 0)									(0, 0.5)								
Comparison groups		2 vs 1			3 vs 1			3 vs 2			2 vs 1			3 vs 1			3 vs 2		
True ATE		0			0			0			0			0.5			0.5		
		EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE	EST	Emp.SE	SE
IPW	Mul	-0.102	0.052	0.082	-0.098	0.05	0.078	0.004	0.022	0.059	-0.1	0.05	0.082	0.404	0.049	0.078	0.503	0.023	0.058
	RF	0.016	0.035	0.065	0.031	0.033	0.062	0.015	0.033	0.06	0.019	0.035	0.065	0.531	0.033	0.062	0.513	0.032	0.06
	GBM	0.034	0.033	0.065	0.041	0.033	0.062	0.007	0.029	0.059	0.036	0.034	0.064	0.541	0.033	0.061	0.506	0.029	0.059
	CBPS	0.048	0.031	0.068	0.039	0.031	0.064	-0.01	0.03	0.057	0.048	0.032	0.068	0.54	0.031	0.064	0.492	0.031	0.057
	MinMean	-0.004	0.065	0.07	-0.001	0.063	0.066	0.003	0.03	0.058	-0.001	0.063	0.069	0.501	0.062	0.066	0.502	0.029	0.058
	MinMax	0.005	0.06	0.069	0.008	0.058	0.066	0.003	0.03	0.058	0.006	0.06	0.069	0.508	0.059	0.065	0.502	0.029	0.058
DR	Mul	0	0.01	0.01	0	0.01	0.01	0	0.008	0.008	0	0.01	0.01	0.5	0.01	0.01	0.5	0.009	0.008
	RF	0	0.01	0.008	0	0.009	0.008	0	0.008	0.008	0	0.009	0.008	0.5	0.009	0.008	0.5	0.009	0.008
	GBM	0	0.01	0.007	0	0.009	0.006	0	0.008	0.006	0	0.009	0.007	0.5	0.009	0.006	0.5	0.009	0.006
	CBPS	0	0.009	0.008	0	0.009	0.008	0	0.008	0.008	0	0.009	0.008	0.5	0.009	0.008	0.5	0.009	0.008
	MinMean	0	0.01	0.009	0	0.009	0.008	0	0.008	0.008	0	0.009	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	MinMax	0	0.01	0.009	0	0.009	0.008	0	0.008	0.008	0	0.009	0.009	0.5	0.009	0.008	0.5	0.009	0.008
enDR	Mul	0	0.01	0.01	0	0.01	0.01	0	0.008	0.008	0	0.01	0.01	0.5	0.01	0.01	0.5	0.009	0.008
	RF	0	0.01	0.008	0	0.009	0.008	0	0.008	0.008	0	0.009	0.008	0.5	0.009	0.008	0.5	0.009	0.008
	GBM	0	0.01	0.007	0	0.009	0.006	0	0.008	0.006	0	0.009	0.007	0.5	0.009	0.006	0.5	0.009	0.006
	CBPS	0	0.009	0.008	0	0.009	0.008	0	0.008	0.008	0	0.009	0.008	0.5	0.009	0.008	0.5	0.009	0.008
	MinMean	0	0.01	0.009	0	0.009	0.008	0	0.008	0.008	0	0.009	0.009	0.5	0.009	0.008	0.5	0.009	0.008
	MinMax	0	0.01	0.009	0	0.009	0.008	0	0.008	0.008	0	0.009	0.009	0.5	0.009	0.008	0.5	0.009	0.008
enOM		0	0.009	0.001	0	0.009	0.001	0	0.008	0.001	0	0.009	0.001	0.5	0.009	0.001	0.5	0.009	0.001

Figure A1.3: Simulation results for Scenario BA (i.e., GPS_B and Out_A), where EST and SE are, respectively, the average of 1000 estimated ATEs and their standard errors. Emp.SE is the standard deviation of the 1000 estimated ATEs.

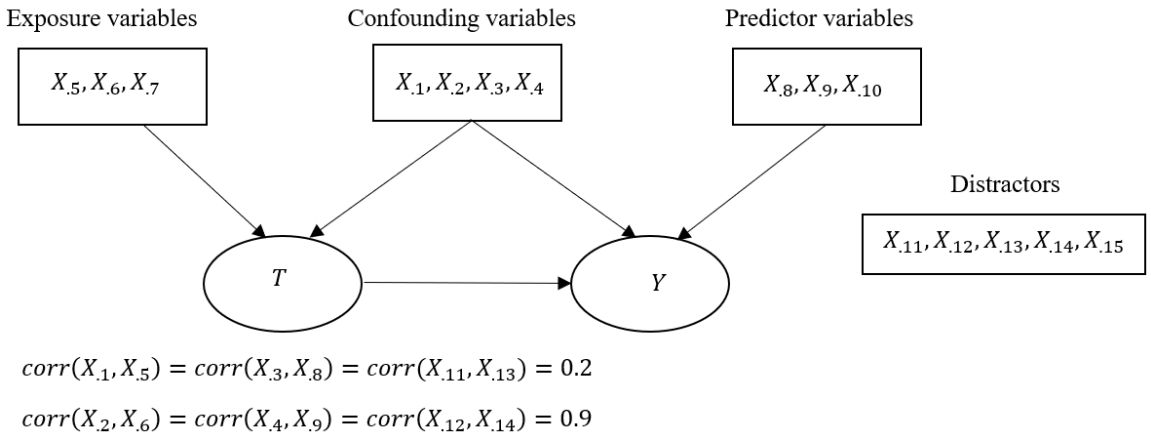

Figure A1.4: Graphic illustration for different types of variables used in the simulation studies in Section 2.4.1.

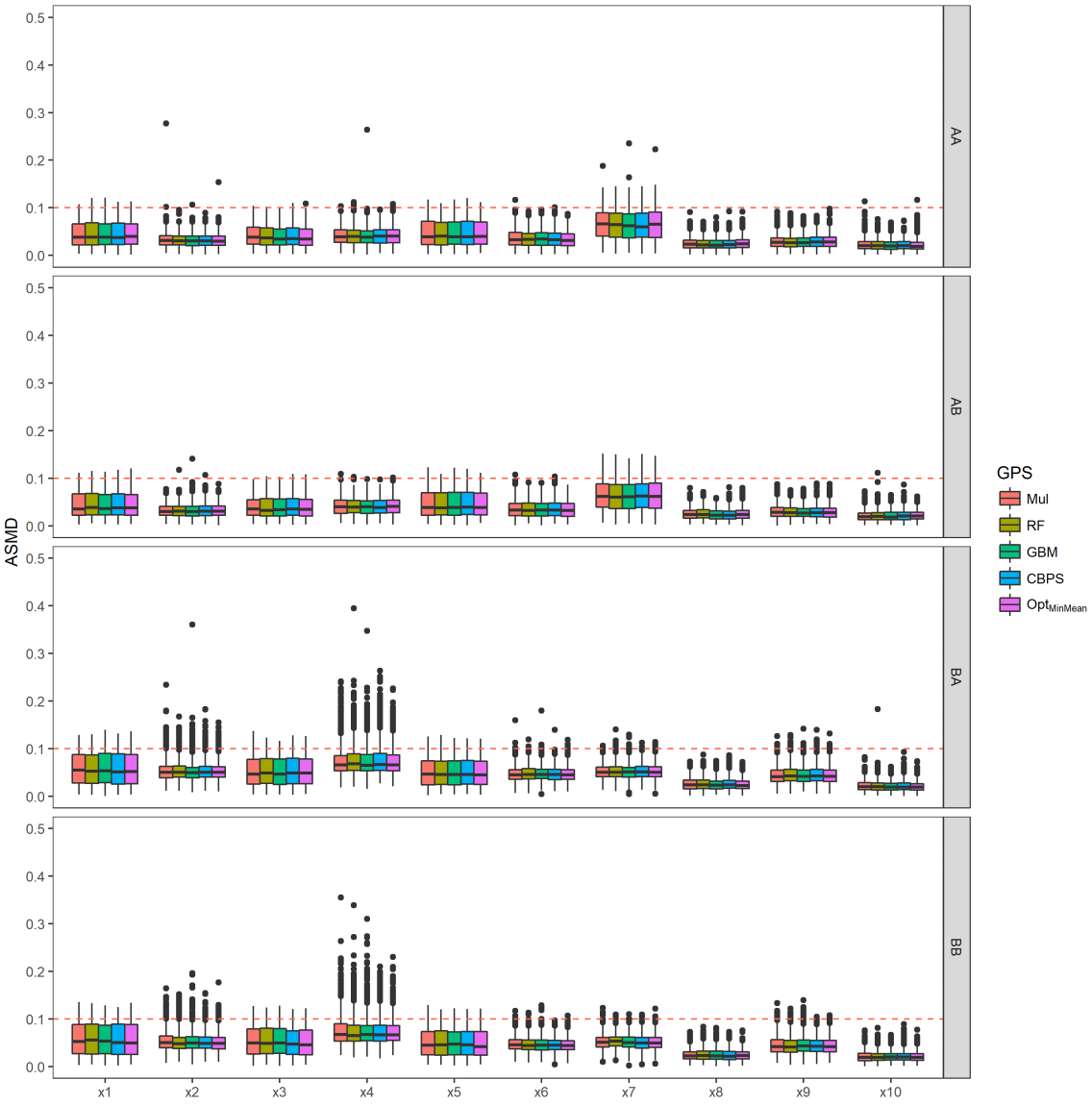

Figure A1.5: The boxplots of 1000 absolute standardized mean differences (ASMDs) based on MinMean criteria for four simulation scenarios under five different GPS estimation methods: multinomial logistic regression (Mul), random forest (RF), GBM, and the covariate balancing propensity score (CBPS), and the optimal GPS estimation method based on MinMean criteria(Opt_{MinMax}),where a lower ASMD indicates a better balance of the covariates.

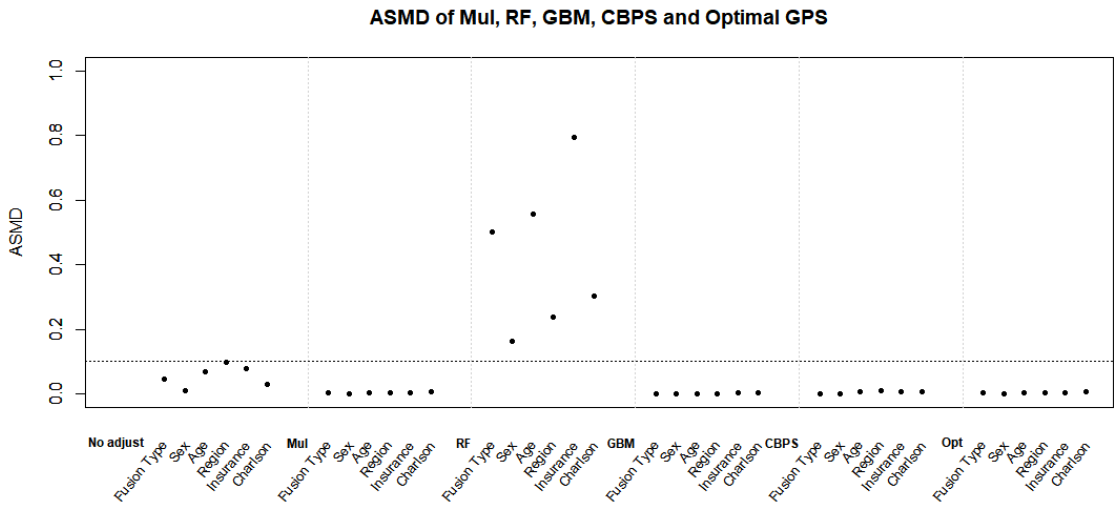

Figure A1.6: Absolute standardized mean differences (ASMDs) for the MarketScan dataset: ASMD without any adjustment (No adjust), ASMDs under four different GPS estimation methods (i.e., multinomial logistic regression (Mul), random forest (RF), GBM, CBPS) and the optimal GPS estimation method based on MinMean criteria (Opt). The covariates (fusion type, sex, age, region, insurance and Charlson comorbidity index) are included in the GPS and outcome model. The horizontal line for $h = 0.1$ is the recommended cut-point on whether a covariate is balanced or not. A lower ASMD indicates a better balance of covariate.

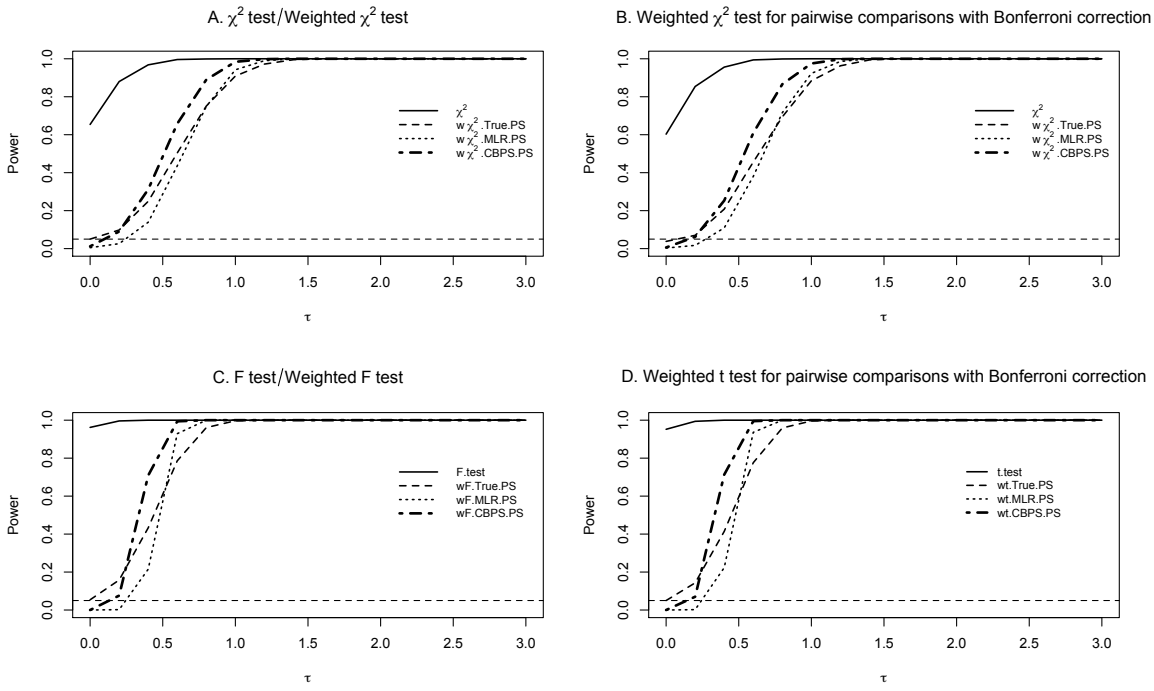

Figure A1.7: Power curves of different tests with sample size 1000. In each panel, the solid line represents the traditional test, the dashed line represents the weighted test using the true GPS, the dotted line represents the weighted test using GPS estimated by multinomial logistic regression (MLR) model, and the dash-dotted line represents the weighted test with GPS estimated using CBPS method. The horizontal line is at a height 0.05, the nominal size of the test.

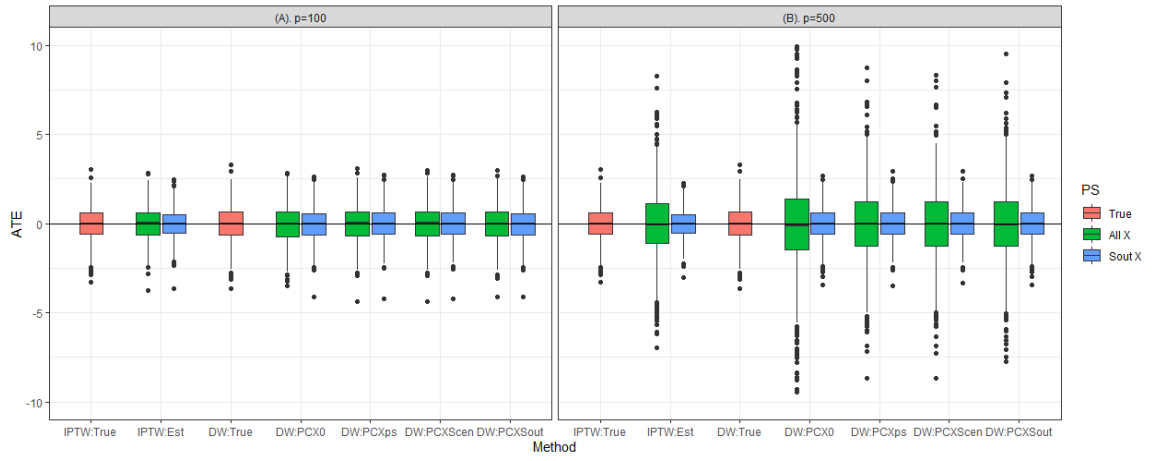

Figure A1.8: The boxplots of 1000 ATE estimates based on IPTW and DW methods, combination with different sets of covariates in the propensity score model, and different sets of covariates in the probability of uncensoring model, under Scenario I.

Table A1.1: Bias and standard error (S.E.) of ATE estimates based on IPTW and DW methods under Scenario I: Non-informative censoring.

	p=100		p=500	
	Bias	S.E.	Bias	S.E.
True ATE	-0.0003	0.0149		
$\boldsymbol{IPTW}$:$True$	-0.0028	0.0285		
$\boldsymbol{IPTW}$:Est				
$PS_{(Conf)}$	-0.0021	0.0279		
$PS_{(Out)}$	0.0044	0.0255		
$PS_{(All)}$	-0.0041	0.0278	-0.0051	0.0626
$PS_{(S_{out})}$	-0.0042	0.0254	-0.0015	0.025
$\boldsymbol{DW}$:$True$	-0.002	0.0314		
$\boldsymbol{DW}$:PC_{X_0}				
$PS_{(Conf)}$	-0.0008	0.0307		
$PS_{(Out)}$	0.0063	0.0028		
$PS_{(All)}$	-0.0059	0.0308	-0.0098	0.068
$PS_{(S_{out})}$	-0.0027	0.028	-0.0013	0.0277
$\boldsymbol{DW}$:$PC_{X_{ps}}$				
$PS_{(Conf)}$	-0.0033	0.0304		
$PS_{(Out)}$	0.0044	0.0278		
$PS_{(All)}$	-0.0151	0.0311	0.0462	0.1226
$PS_{(S_{out})}$	-0.0111	0.0277	-0.0022	0.0275
$\boldsymbol{DW}$:$PC_{X_{S_{cen}}}$				
$PS_{(Conf)}$	-0.0035	0.0306	-0.0006	0.0306
$PS_{(Out)}$	0.0034	0.0281	0.0062	0.0281
$PS_{(All)}$	-0.0077	0.0308	-0.0159	0.068
$PS_{(S_{out})}$	-0.0056	0.028	-0.0018	0.0277
$\boldsymbol{DW}$: $PC_{X_{S_{out}}}$				
$PS_{(Conf)}$	-0.009	0.0305	-0.001	0.0306
$PS_{(Out)}$	-0.0017	0.0278	0.0062	0.0279
$PS_{(All)}$	-0.0165	0.0306	-0.0039	0.0691
$PS_{(S_{out})}$	-0.0111	0.0277	-0.0022	0.0275

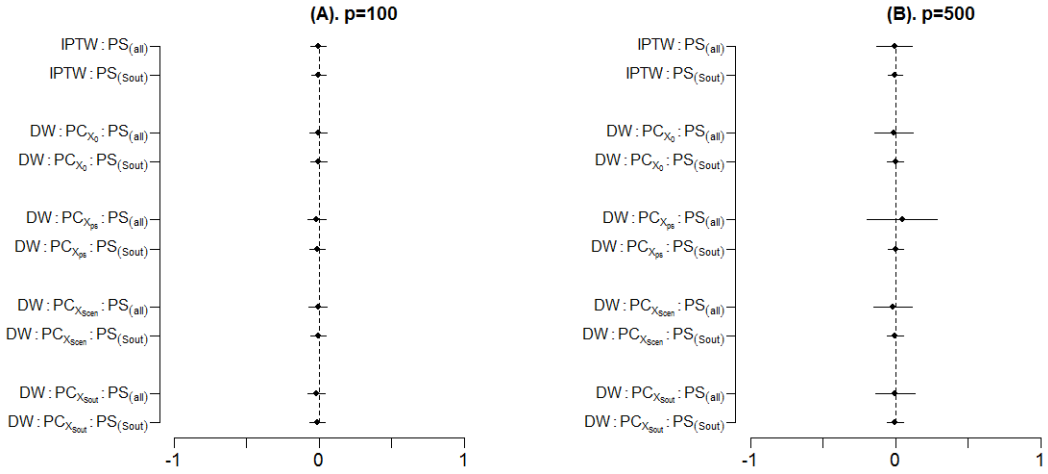

Figure A1.9: ATE and their 95% CI of estimates for p=100 and 500 under Scenario I.

CURRICULUM VITA

NAME: Xiaofang Yan

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40292

EDUCATION: Bachelor of Science in Mathematics,
Xi'an Jiaotong University, 2011
Masters in Probability and Statistics,
Chinese Academy of Sciences, 2015

PUBLICATIONS: Yan, X., Zheng, Q., Kong, M. Estimation of treatment effect for time-to-event outcomes. (Under preparation)

Yan, X., Zheng, Q., Kong, M. Weighted χ^2 and F test for multiple group comparisons in observational studies. (Under review)

Yan, X., Abdia, Y., Datta, S., Kulasekera, KB, Ugiliweneza, B., Boakye, M., Kong, M (2019). Estimatcion of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in Medicine*, 38(15):2828-2846.

Li, B., Hao, J., Yan, X., Kong M., Sauter, E (2019). A-FABP and estrogens are independently involved in the development of breast cancer. *Adipocyte*. (Accepted)

Young, J., Yan, X., Xu, J., Yin, X., Zhang, X., Arteel, G., Barnes, G., States, J., Watson, W., Kong, M., Freedman, J., Cai, L (2019). Cadmium and high-fat diet dispute renal, cardiac and hepatic ssential metals. *Scientific Reports*. (Accepted)

Davis, D., Feygin, Y, Creel, L., Williams, P., Lohr, W., Jones, V., Le, J., Pasquenza, N., Ghosal, S., Jawad, K., Yan, X., Liu, G., McKinley, S (2018). Longitudinal trends in the diagnosis of attention-deficit/hyperactivity disorder and stimulant use in preschool children on Medicaid. *The Journal of Pediatrics*, Doi: 10.1016/j.jpeds.2018.10.062.

Hao, J., Zheng, Y., Yan, X., Kong, M., Li, B., et al (2018). Circulating adipose fatty acid binding protein is a new link underlying obesity-associated breast/mammary tumor development. *Cell Metabolism*, 28(5):689-705.

Kolluru, V., Chandrasekaran, B., Tyagi, A., Dervishi, A., Ankem, M., Yan, X., Damodaran, C., et al (2017). miR-301a expression: diagnostic and prognostic marker for prostate cancer. *Urologic Oncology*, 36(11):503-e9.

Li, B., Hao, J., Yan, X., Kong, M., Sauter, E (2017). A-FABPdecreases in the wean milk of nursing women with a family history of breast cancer. *International Journal of Women's Health and Wellness*, 3(063):2474-1353.

PRESENTATIONS: Guest lecturer for ggplot2 package, University of Louisville, KY, September 2019.

Estimation of average treatment effects among multiple treatment groups by using an ensemble approach, ASA-KY Chapter Meeting, Louisville, KY, March 2019.

Weighted F and χ^2 test statistics for testing treatment effect among multiple groups in observational studies, ASA-KY Chapter Meeting, Louisville, KY, March 2018.

HONORS AND AWARDS

Harshbarger Travel Award, Southern Regional Council on Statistics (SRCOS). June 2019

Best Student Presentation Award, ASA-KY Chapter Meeting. March 2018

Graduate Fellowship, University of Louisville. August 2015 -May 2017